UNIVERSIDADE D
COIMBRA

Ana Teresa Amado Mateus Santos Rajado

Computation meets experimentation to improve the catalysis and specificity of Cas12a genome editing enzyme

Dissertação no âmbito do Mestrado em Bioquímica orientada pela Doutora Alexandra Teresa Pires Carvalho e pela Senhora Professora Paula Cristina Veríssimo Pires e apresentada ao Departamento de Ciências da Vida.

Novembro de 2020

Computation meets experimentation to improve the catalysis and specificity of Cas12a genome editing enzyme

Ana Teresa Amado Mateus Santos Rajado
Master's Degree in Biochemistry
Faculty of Science and Technology
University of Coimbra

Supervisors:
Alexandra Carvalho, PhD (UC-Biotech)
Paula Veríssimo, PhD (Department of Life-Sciences)

"If the eternal dance of molecules
Is too entangled for us mortal fools
To follow, on what grounds should we complain?
Who promised us that Nature's arcane rules
Would make sense to a merely human brain?"
Peter Shore

The work presented on this dissertation was developed with the “Rational Protein Engineering” Group at UC-Biotech - Center for Neuroscience and Cell Biology, with funding from FCT (grant IF/01272/2015) and the experimental work was developed with the Structural Biotechnology Group at UC-Biotech - Center for Neuroscience and Cell Biology. This work was also funded by FCT and DGES under the program PEES.

Agradecimentos

Chegando ao culminar do que acredito que será uma das etapas mais importantes da minha vida, não a poderia deixar terminar sem agradecer a algumas pessoas que de uma forma ou de outra contribuíram para o meu crescimento e para que me tornasse quem hoje sou.

À Doutora Alexandra Carvalho, quero agradecer por me ter aceite como orientanda, abrindo-me assim as portas para um mundo novo. Sem a oportunidade que me deu, talvez nunca viesse a descobrir este caminho e passasse ao lado de uma área extremamente rica sem pestanejar, por ter o "preconceito" de que "computadores não são para mim". Estou imensamente grata por todo o apoio, sugestões, explicações dadas e disponibilidade constante, sem as quais certamente não teria chegado onde cheguei.

Ao Doutor Ricardo Vieira-Pires, quero agradecer a receção no laboratório, tantas vezes a contra relógio, sem nunca desmotivar e encontrando sempre um bocadinho para me ajudar, alimentando o meu espírito crítico e motivando-me a superar-me a mim mesma.

À Doutora Paula Veríssimo, gostaria de agradecer ter aceite ser a minha orientadora interna e, mais que isso, pelas chamadas à realidade, que me ajudaram a não desmotivar quando as etapas que rigidamente me propus a cumprir dentro de prazos muito concretos nem sempre coincidiam com aquilo de que estava à espera.

Ao Grupo de Investigação em que fui acolhida - em especial à Beatriz, à Sónia, ao Pedro e ao Daniel -, um muito obrigada por tudo o que me ensinaram, tanto a nível pessoal como profissional, pelos conselhos, brincadeiras e paciência, que tornaram esta experiência muito mais enriquecedora do que estava à espera quando a ela me propus.

Ao Doutor Warispreet Singh quero agradecer todo o apoio que me deu ao longo deste projeto, tal como toda a disponibilidade e amabilidade.

Não podia deixar de agradecer aos meus queridos amigos, incansáveis e inabaláveis em todos os momentos. À Catarina, pela "motivação diária" ao longo desta tese e por todo o apoio, motivação e conselhos nos dramas - mais reais ou mais imaginários - que vivi ao longo da última década, bendita Geografia. À Joana, que nem imaginava no que se estava a meter quando me aceitou como afilhada, e que aturou risos e choros sem vacilar. À Patrícia e ao Hugo, pela boa disposição constante, por todos os jogos jogados, jantares jantados e risos partilhados. À Margarida e à Rita, pelas brincadeiras e conversas, pela massa com atum, e por todos os encontros que esperávamos ser fugazes mas que invariavelmente acabavam por se estender.

Quero agradecer à TAUC pelos escapes e à Doutora Sónia por tudo o que me ensinou.

Ao Miguel, para quem todas as palavras são poucas, quero agradecer por ter sido uma rocha no meio da tempestade e por me devolver à realidade quando os sonhos me queriam arrastar para longe. Um Muito Obrigada por todo o carinho, por me relembrar o caminho quando me rendo às dúvidas e pela dedicação demonstrada dia após dia.

E por fim, à minha família, em especial aos meus pais e ao Pedro, muitíssimo obrigada por tornarem tudo isto possível, porque sem vocês esta história, que é a minha vida, teria tido uma narrativa muito diferente. Devo-vos o mundo, pois foi o mundo que me deram, com todas as ferramentas que me foram facultando ao longo destes 23 anos de vida, entre valores, aprendizagens e amor - sempre com apoio incondicional e muitas vezes com mais confiança que eu nas minhas próprias capacidade, sempre na fila da frente a celebrar entusiasticamente as minhas vitórias ou a relembrar-me que as dificuldades são apenas desafios à espera de ser superados.

A todos aqueles que marcaram o meu percurso, entre professores, funcionários, colegas e amigos, o meu muito obrigada.

Abstract

The CRISPR-Cas system is a tool used for genome editing that became highly relevant in the latest years for being cheap, easy to design and produce.

Cas12a is an endonuclease type V of the CRISPR-Cas system and is able to edit human genome through a single-RNA guided approach. This enzyme has already been repurposed to be applied in several fields, such as in medicine and agriculture, through the genome editing of different cells types such animal and plant cell. However a recurrent problem of these systems and related ones is the off-target mutations - unintentionally induced.

The objective of this work is to study Cas12a enzyme. For this, we used a combination of computational (Molecular Dynamics) and experimental (Molecular Biology) methods, in order to surpass the above mentioned obstacles.

Six variants of the wild type enzyme (directed to enzyme' regions that interact with the PAM motif, the crRNA 5' handle and with the active site of the protein) and two putative intermediates of its catalytic cycle were tested. With these variants, we induced stronger interactions between FnCas12a, a genome editing enzyme, and its crRNA and target DNA.

Additionally, we explored the catalytic mechanism of this enzyme, by studying the relevance of H922 and R1218, residues located in the catalytic site of the enzyme.

Keywords

CRISPR-Cas System; Cas12a; Single Guide RNA; Genome Editing; Computational Biology; Rational Protein Engineering

Sumário

O sistema CRISPR-Cas é uma ferramenta aplicada a edição genética. Esta técnica tornou-se altamente relevante nos últimos anos devido ao seu baixo custo e facilidade de produção e utilização.

Cas12a é uma endonuclease do tipo V do Sistema CRISPR-Cas, capaz de editar genoma humano recorrendo a um único RNA guia. Esta enzima já foi adaptada e utilizada em diversas áreas, tal como medicina e agricultura, através da edição genética de células de diferentes tipos, como por exemplo células animais e vegetais. No entanto, este sistema também enfrenta alguns problemas, dos quais se destacam as mutações introduzidas fora dos locais alvo ("off-target mutations"), que são introduzidas de forma não intencional.

O objetivo deste trabalho é estudar o mecanismo catalítico da enzima Cas12a, com o intuito de aumentar a especificidade do mesmo. Com este propósito recorremos a uma combinação de métodos computacionais (Dinâmica Molecular) e experimentais (Biologia Molecular), para reduzir os efeitos "off-target" acima mencionados.

Foram estudadas seis variantes da enzima nativa (direcionadas para as regiões da enzima que interagem com o motivo PAM, com o loop no terminal 5' do crRNA e com o centro ativo da enzima) e dois estados intermediários do ciclo catalítico da mesma. Com as variantes criadas induzimos interações mais fortes do que as previamente presentes entre a FnCas12a, uma enzima para edição genética, e o crRNA e DNA alvo a ela associados.

Explorámos também o mecanismo catalítico desta enzima, ao estudarmos a relevância dos resíduos H922 e R1218, localizados no local catalítico da enzima.

Palavras chave

Sistema CRISPR-Cas; Cas12a; RNA Guia Único; Edição Genética; Biologia Computacional; Engenharia Racional de Proteínas

Index

Abbreviations

This list contains the abbreviations within this work. Every measure unit is used in accordance to the international unit system. The nomenclature for aminoacids is presented in accordance to the three-letter code - where the first three letters of the aminoacid name are the abbreviation - and to the one-letter code, devised by Margaret Oakley Dayhof. The nomenclature for nucleic acids is in accordance to the AMBER libraries. In these, the DNA nucleotides have a "D" before the letter which designates the nucleotide that is being referred (adenine - DA -, cytosine - DC -, guanine - DG - and thymine - DT), while the RNA nucleotides are only designated by a single letter (adenine - A -, cytosine - C -, guanine - G -, and uracil - U).

AMBER - Assisted Model Building with Energy Refinement
CRISPR - Clustered Regularly Interspaced Short Palindromic Repeats
crRNA - CRISPR RNA
DFT - Density Functional Theory
gRNA - guide RNA
sgRNA - single-guide RNA
DNA - Desoxiribonucleic Acid
MD - Molecular Dynamics
MM - Molecular Mechanics
PMEMD - Particle Mesh Ewald Molecular Dynamics
QM - Quantum Mechanics
QM/MM - Quantum Mechanics/Molecular Mechanics
RNA - Ribonucleic Acid
TALEN - Transcription Activator-Like Effector Nucleases
VMD - Visual Molecular Dynamics
WT - Wild Type
ZFN - Zinc Finger Nucleases

Figure Index:

Table Index

Equations Index

Objectives

The present work aims to study the Cas12a enzyme *in silico* and *in vitro*, to gain further insight on the dynamical nature of the interactions between the enzyme and the nucleic acids (DNA and crRNA). Ultimately, the *in silico* experiments and simulations will potentially help to understand the catalytic mechanism of Cas12a and improve the specificity of the manipulated enzyme by increasing the number of interactions between the protein and the nucleic acids, promoting base specific recognition.

To achieve these objectives we propose to perform: (i) Molecular Modelling to optimize a chosen Cas12a structure; (ii) Molecular Mechanics, applied through Molecular Dynamic simulations, of the Wild-Type (WT) structure and with basis on the WT MD simulations (iii) study enzymatic variants that should aid the comprehension of this protein catalysis, as well as mutants which will target nucleic acid regions proved to be relevant for Cas12a activity; (iv) Quantum Mechanics/Molecular Mechanics to characterize the catalytic mechanism of the enzyme; (v) In vitro validation of the most promising mutant to increase the enzyme specificity.

The last step proposed was compromised by Sars-CoV-2 pandemic and the mandatory quarantine. As a result, it was only possible to reach the experimental step regarding the expression of the wild type enzyme, during the protocol optimization, and it was not possible to proceed to the validation of the selected mutant. The obtained results and the work performed - as well as the projected steps to be developed - will be further explained in the chapters ahead.

Introduction

Protein engineering:

Proteins are macromolecules composed by a linear sequence (ranging from a few dozens to several thousand residues in length (Purves, *et al*., 2004)) of the same 20 amino acids existent in every living organism, distributed through countless combinations (Vieira, *et al*., 2012) and structural organizations (Fuxreiter, 2015). These polymers are likely the most versatile of all biomolecules (Nelson, 2005), being responsible for several functions within any form of life, including: structural support, protection, transport, enzyme catalysis, defence, regulation, signal transmission and movement (Purves, *et al*., 2004). Due to their intrinsic versatility, it is possible to find a great variety of proteins in a single cell (Vieira, *et al*., 2012). The residues organization in the sequence allows these molecules to have enough information, promoting the acquisition of their three dimensional (3D) structure (which may be aided, for example by chaperones), which is reflected in said variety of functions (Nelson, 2005).

The structure of a protein can be considered at 4 levels: primary structure - which considers the protein amino acid sequence; secondary structure - which includes helices (such as the alpha-helice), sheets (as the beta-sheet) and loops; tertiary structure - combines the secondary structure elements; and quaternary structure - which describes the way the different subunits are organized (Nelson & Cox, 2008).

The study of enzymes can be traced back to the beginning of the nineteenth century, even though they have always been employed by humankind, for example in the production of beer and cheese. Nonetheless, this class of proteins was only classified as such in 1930 (Aehle, 2007). By 1965, Blake and colleagues, through Fourier synthesis, were the first to obtain the structure of an enzyme with high resolution (2 Å) - the hen egg white lysozyme (Blake, 1965). This breakthrough opened the doors to the "structure-function studies of enzymes" (Carvalho, *et al*., 2014). Due to technological advancements, these studies progressed from merely *in vitro* to *in silico* experiments. In 1975, through computational methods, Levitt and Warshel explored the tertiary folding on bovine pancreatic trypsin inhibitor by combining energy minimization with normal mode thermalisation, restricting the freedom degrees and the number of protein interactions (Levitt & Warshel, 1975).

Protein engineering can be defined as "the design of new enzymes or proteins with new or desirable functions" (Turanli-Yildiz, *et al*., 2012) and it is considered an important tool to overcome some enzymes' limitations (Arab, *et al*., 2014). This field of study owes its prompt growth to the developments observed in biosciences, especially recombinant DNA technology. Thus, it is now possible to resort to several methods to study and modify proteins, from random to semi-rational and rational approaches, both experimentally and computationally (Harris & Craik, 1998; Turanli-Yildiz, *et al*., 2012). A representation of rational and random approaches is presented in the figure below. Rational approaches start with the study of a protein structure and derive from there, trying to improve determined characteristics of a protein through a defined plan and with theoretical knowledge of the molecule beforehand. On the other hand, random approaches generate random mutations and the variants of interest are selected afterwards from a pool. Both procedures may be repeated as many times as necessary until achieving a mutant with the intended characteristics (Bornscheuer & Pohl, 2001).

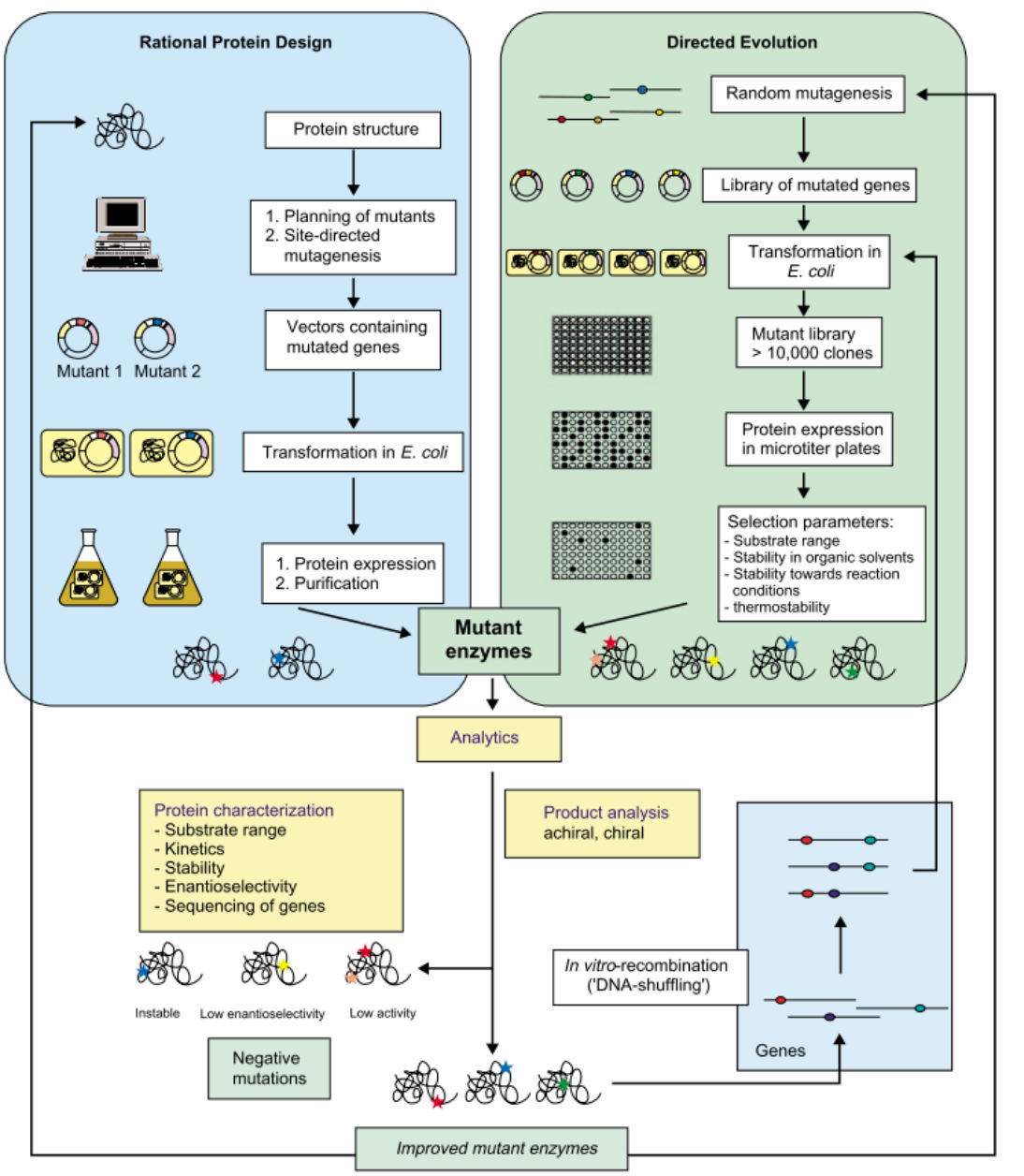


Figure1 - Representation of rational and random approaches to protein design (Bornscheuer & Pohl, 2001).

Computational chemistry

In the present work, computational chemistry methods are applied to predict Cas12a structures at different steps of the catalytic mechanism and to understand the impact of the selected mutations.

We can divide computational design in protein or enzyme redesign and *de novo* protein design. Both aim at the creation of "artificial" proteins, specially built to fit the existing needs, when the existent ones do not have the ideal catalytic activity or binding interface,

by characterizing the 3D structure of a protein and identifying or introducing properties of interest (Lippow & Tidor, 2007; Wijma & Janssen, 2013). The conception of these molecules became possible thanks to *in silico* procedures (and the increasingly more reliable computational modelling methods used), the existent knowledge about polymers and empirical search in specialized libraries (Turanli-Yildiz, *et al*., 2012; Bogdanove, *et al*., 2018). The new designed molecules - mutants - are ranked and if the predictions are not good, the calculations are usually repeated with different parameters (Wijma & Janssen, 2013).

This work has as focus protein redesign. This can be used to alter an enzyme's specificity, thermostability, binding, catalysis and affinity to different substrates (Lippow & Tidor, 2007).

Computational chemistry techniques may be divided in: Quantum Mechanical (QM), based in wave functions, among which may be found ab initio and semi empirical methods, as well as methods based in the Density Functional Theory (DFT). QM methods are very costly approaches, from a computational point of view, and cannot be applied to big systems (e.g. DFT can accurately describe about 100 atoms), however, these methods allow a very precise mathematical description of an atomic system (Ramos, 2012; Bakoweis & Thiel, 1996); (ii) methods with base on Classical Mechanics, which are empirical, as they make use of parameters fitted to experimental data or obtained from QM calculations and described in force-fields to study a molecule mechanism (Ramos, 2012; The Sherrill Group, n.d.) and (iii) hybrid methods which combine Quantum Mechanics with Molecular Mechanics (QM/MM), using QM to define the protein' catalytic site with higher precision - having in attention the inter-atomic interactions and surface energy values - allowing the attaining of a detailed comprehension of a molecule's behaviour during its catalytic activity, and MM, a less costly (from a computational perspective) methodology (Carvalho, *et al*., 2014; Melo, *et al*., 2018), to model the remaining residues after establishing their adequate constraints, interactions and boundaries (Friesner & Guallar, 2005) and to include the protein environment (Ramos, 2012, Carvalho, *et al*., 2014).

Molecular Dynamic (MD) simulations is a computational methodology that aims to study the movement of the molecular components of a system with a microscopic model (Vieira, *et al*., 2012). It resorts to different techniques and algorithms to better understand and predict the molecule's behaviour in a determined context, in accordance to Newton's laws of motion (Leach, 2001). In classical molecular dynamics, the molecular systems

are represented by different spheres connected by springs, described by a simple potential energy function and corresponding parameters. It is important to contextualize the molecule atoms in the whole system - being pertinent to specify the spheres' position within the molecule structure (Leach, 2001).

Protein engineering has increased in popularity in the last years, not only due to its scientific value, but also because of its applications (Brannigan & Wilkinson, 2002; Turanli-Yildiz, *et al*., 2012). This technology evolution has led to the development of enzymes introduced in several fields, some examples are: in agriculture, protein engineering lead to the introduction of a modified thioesterase in crop plants to "increase the proportion of unsaturated fatty acids in seed oil"; in medicine, where it allowed the production of recombinant proteins, such as insulin or antibodies, for therapeutic purposes; in the food industry, where it promoted the development of new ingredients as well as food-processing enzymes and flavours adjustment enzymes; in the detergent industry, where designed proteases and lipases were integrated in products to remove stains; in environmental care, where it is applied to bioremediation and to the development of biosensors and recently to the development of bioplastics (Facciotti, et al., 1999; Brannigan & Wilkinson, 2002; Turanli-Yildiz, et al., 2012).

Genome editing:

Genome editing dates back to the 1970s with the development of recombinant DNA technology. It is a tool based in genome engineering (Hsu, 2014), allowing the induction of mutations to specific targets within the genome by using programmable nucleases (Zhang, *et al*., 2014), and was a milestone in the history of biosciences, due to the possibilities it provided (Hsu, 2014) (Figure 2). This technology is employed to study functional genomics, transgenic organisms and gene therapies (Zhang, *et al*., 2014), as well as the development of genetically modified organisms, drugs and therapies, synthetic materials, fuels and food, to give some examples (Hsu, 2014).

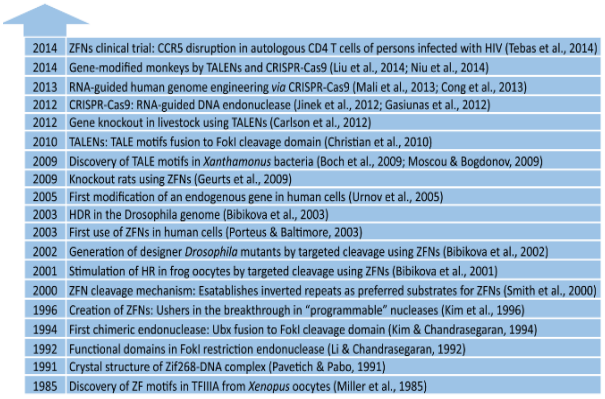


Figure 2 - Milestones in genome editing (Chandrasegaran & Carroll, 2016).

A better understanding of gene functions is achieved through the comparison of a wild-type variant and a system with a modified or even deleted gene (Chandrasegaran & Carroll, 2016). Gene editing is possible because of the cell mechanisms for DNA repair (Zhang, *et al*., 2014), which can be divided in two classes: non-homologous end joining and homology-directed repair, which will be used to re-join the DNA strands after a double-break induced by nucleases (Chandrasegaran & Carroll, 2016).

The possibility to edit an individual's genome exists due to cell mechanisms for DNA repair. After the DNA double strand break is induced, sequences codifying to the acquisition of properties of interest will be introduced in the system - as homologous templates - what may result in their integration in the genome (Chandrasegaran & Carroll, 2016) (Figure 3).

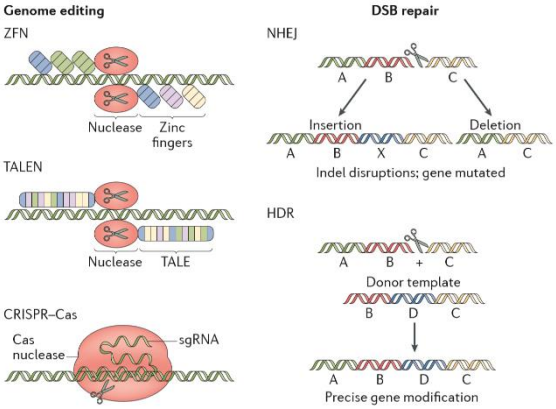


Figure 3 - Mechanisms of genome editing (able to induce double strand breaks (DBS)): Zinc Finger nuclease (ZFN), Transcription Activator-Like Effector Nuclease (TALEN) and Clustered Regularly Interspaced Short Palindromic Repeats and its associated protein (CRISPR–Cas system) and DBS repair mechanisms (used to achieve genome editing): non-homologous end joining (NHEJ) and homology-directed repair (HDR) (Yin, *et al.*, 2017).

Zinc Finger Nucleases

The first class of programmable nucleases used for genome editing were Zinc Finger nucleases (ZFN). These proteins are constituted by two main domains (Chandrasegaran & Carroll, 2016): a Cys_2-His_2 domain, that binds to DNA and a domain responsible for cleavage of the DNA - FokI restriction endonuclease (Zhang, *et al*., 2014). Cys_2His_2 zinc fingers are the most common motifs of DNA-binding protein domains - although they can also recognize and bind RNA and proteins. These structures are usually short in length (approximately 30 amino acid) and are composed by a zinc ion coordinated to two cysteines and two histidine residues (Razin, *et al.*, 2012). It was demonstrated that inducing a DNA double strand break with ZFN promoted DNA repair by homologous recombination. In most cases, the zinc finger motifs have the capacity to recognize and bind three base pairs of the sequence. To increase the number of base pairs recognized, and consequentially improve the specificity of the enzyme, chimeras with several motifs in a row projected to the sequence of interest in the target DNA were created

(Chandrasegaran & Carroll, 2016). However, the capacity of these altered nucleases to induce specific DNA breaks was lower (Yin, *et al*., 2017). Nonetheless, Urnov and colleagues used this mechanism to correct, for the first time in human cells, a gene mutation disease inducing (Zhang, *et al*., 2014).

Transcription Activator-Like Effector Nucleases

Another tool for genome editing is Transcription Activator-Like Effector Nucleases (TALEN) (Zhang, *et al*., 2014). TALENs, similarly to ZFNs, are able to recognize a DNA target through protein-DNA interactions (Yin, *et al*., 2017) and owe their cleavage capacity to a FokI nuclease. In opposition to the first method presented, TALENs have a higher capacity to induce double strand breaks and mutagenesis (Yin, *et al*., 2017). They are also composed by several repeats of 33 to 35 amino acid residues that vary from each other at the positions 12 and 13. These positions specify which nucleotide of the target DNA is recognized (since every unit of 33-35 residues is only able to recognize one nucleotide) (Zhang, *et al*., 2014; Chandrasegaran & Carroll, 2016). Therefore, just as it happens with ZFNs, the association between several of the referred recognition units, allows the recognition of a longer DNA target sequence, which may induce specific DNA breaks (Chandrasegaran & Carroll, 2016).

CRISPR

The last tool for genome editing presented in this work, and the one that will be explored herein, is CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats). This tool differs from ZFN and TALEN because it is not guided by a protein-DNA interaction, instead, it depends on DNA-RNA recognition and only needs a small guide RNA (gRNA) for target identification. This makes this technique cheaper and easier to design, produce and apply than the previously mentioned tools (Zhang, *et al*., 2014; Chandrasegaran & Carroll, 2016; Hille & Charpentier, 2016; Zhang, *et al.*, 2017).

CRISPR and its associated genes/proteins (Cas) form a mechanism able to confer immunity to bacteria and archaea against virus and other invasive genetic elements, such as plasmids (Heler, *et al*, 2014; Swarts & Jinek, 2018). By 2016, this mechanism had been observed in approximately 40% of the bacterial genomes sequenced and in about 90% of the analysed genomes from archaea (Zhang, *et al*., 2014). The system evolved in time and acquired the capacity to recognize previous invaders by integrating DNA segments as "spacers" in the CRISPR array (Heler, *et al*, 2014), through two classes of nuclease effectors (Zetsche, *et al*., 2015) divided between six types of CRISPR (Swarts & Jinek, 2018). The effectors belonging to Class 1 use multi-protein complexes to recognize and cleave the DNA, while Class 2 effector nucleases use a single molecule (such as Cas9 or Cpf1) for the same purpose (Zetsche, *et al*., 2016; Hille & Charpentier, 2015).

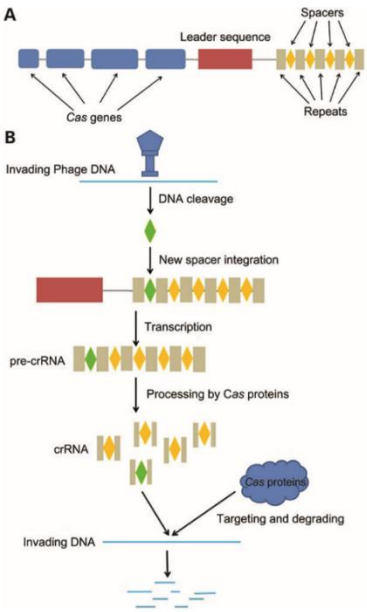


Figure 4 - Representation of CRISPR-Cas system (A) and its mechanism (B) (Zhang, *et al*., 2014).

The CRISPR-Cas system mechanism of defence can be divided in three stages: 1) adaptation; 2) expression; 3) interference (Swarts & Jinek, 2018) (Figure 4):

1) Adaptation is the phase when the invading DNA is identified as a protospacer by the Protospacer Adjacent Motif (PAM) domain and inserted in the CRISPR array as a spacer, between repeats. In one of the most studied CRISPR-Cas systems, CRISPR-Cas9, it is believed that after the protospacers selection, Cas9 is responsible for recruiting Cas1 and Cas2 protein complexes to complete the spacer integration into the prokaryote DNA (Hille & Charpentier, 2016; Zetsche, *et al*., 2015; Swarts & Jinek, 2018). After DNA recognition and binding by Cas proteins, a double strand break is induced in the target DNA, which will prompt the cell DNA repair mechanisms (Figure 2) (Zhang, *et al*., 2014). In Cas9, the HNH domain is responsible for the cleavage of the complementary DNA strand, while the non-complementary strand is cleaved by the Cas9 RuvC-like domain (Zhang, *et al*., 2014; Jinek, *et al.*, 2012).

2) The expression stage occurs when the CRISPR array is transcribed - originating long transcripts (precursors of CRISPR RNA - or pre-crRNA) that mature into crRNAs. These consist of fragments of spacers introduced for specific invaders and a segment of

their adjacent repeats (Hille & Charpentier, 2016; Yamano, *et al*., 2016). In the case of Cas9, the crRNA is combined with transactivating crRNA, forming the guide-RNA (gRNA), a complex that will guide Cas9 to a specific DNA target once the cell contacts again the same invaders (Zhang, *et al*., 2014).

3) During interference, the sequences of the formerly integrated invading genomes are recognized as targets by mature crRNAs and will be silenced through cleavage Cas9 (by HNH and RuvC-like domains, in Cas9s' case, or in Cas12a, RuvC) (Jinek, *et al.*, 2012). In this step, Class 1 CRISPR-Cas uses the complex CASCADE (CRISPR-associated Complex for Antiviral Defense) to degrade the targets, while class 2 systems use a single guide RNA (as referred before) (Hille & Charpentier, 2016). The relevance of the PAM is evident in this stage, since it is due to this short sequence that the cell is able to distinguish its own genetic material from information of a foreign source: only protospacers associated with the PAM recognized by the Cas protein in use are cleaved (Heler, *et al*., 2014).

Genome editing with CRISPR-Cas systems can result in several applications with high biological potential, because of this tool highly adaptive nature. CRISPR-Cas systems are also interesting because of the great variety of roles that they are able to perform, from DNA repair, to gene regulation, passing by genome evolution (among others) (Hille & Charpentier, 2016). However, these techniques also face some daunting challenges, such as: off-target effects; dependence on a specific PAM, specific gRNA design for the sequence of interest and the delivery methods in use (Zhang, *et al.*, 2014).

Off-target mutations are one of the main concerns when editing an eukaryotic system. Curiously, CRISPR-Cas systems have a higher risk of off-target modifications in human cells than ZFNs or TALENs (Zhang, *et al*., 2014), limiting its applications. Strategies are being developed to reduce this problem, one of which is the increase of the nucleases specificity (Aouida, *et al*., 2015; Hille & Charpentier, 2016). Considering the high occurrence of off-target effects, tools were developed, such as Cas-OFFinder, to "search for potential off-target sites (of Cas9 RNA-guided endonucleases) in a given genome or user-defined sequences" (Bae, *et al*., 2014).

Cpf1 or Cas12a:

Cpf1 or Cas12a is a type V (class 2) endonuclease of the CRISPR-Cas system (Yamano, *et al*., 2016), originally from Prevotella and Francisella1 (Zetsche, *et al*., 2015),

which has the peculiarity of being able to edit genome through a single-RNA guided approach (Li, *et al*., 2017). This protein, just as the previously mentioned (and better characterized) Cas9, can be programmed to bind and cleave specific DNA targets. However, Cas12a also presents some differences, making it an alternative to the previously mentioned Cas (Swarts & Jinek, 2018). The first difference is that Cas9 is a type II endonuclease, while Cas12a is an endonuclease type V (Yamano, *et al*., 2016), this means that although Cas9 needs to use homologous RNA-guided endonucleases as effectors, Cas12a is able to mature without the need of a transactivating crRNA (Zetsche, *et al*., 2015) and it only requests a single crRNA as a guide (to cleave its target DNA) (Aouida, *et al*., 2015*;* Zhang, *et al*., 2017). Other significant difference is the fact that Cas9 recognizes a PAM rich in guanine (NGG - in which "N" can represent any nucleotide) in the protospacers' 3' terminal (Zhang, *et al*., 2017), while Cas12a prefers the use of a thymine rich motif (TTTN) (Figure 4E) (Kleinstiver, *et al.*, 2016). Furthermore, while Cas9 induces double strand breaks with the creation of blunt ends near the PAM site, Cas12a generates a staggered cut, at 70 Å from PAM's 5'-terminal (Gao, *et al*., 2016; Stella, *et al*., 2018). The staggered double strand break produced by Cas12a results from the different paths followed by the target and by the non-target strand of the genome intended to edit (Stella, *et al*., 2018). These differences between Cas9 and Cas12a are summarized in Table 1.

Table 1 - Representation of some of the main differences between Cas9 and Cas12a (Swarts and Jinek, 2017; Kleinstiver, *et al*., 2016).

Characteristics	**Cas9**	**Cas12a**
Type	II	V
PAM	Guanine rich - NGG	Thymine rich - TTTN
Double Strand Break	Blunt ends	Staggered ends
Cleavage - region	Upstream the PAM	Downstream the PAM
Cleavage - responsible domains	Template strand - HNH Non-template strand - RuvC	Template strand and Non-template strand - RuvC
Associated RNA	crRNA and trRNA	crRNA
Complementarity with DNA protospacer region	20 nucleotides	23 nucleotides

Nonetheless, it is difficult to compare Cas9 and Cas12a genome-wide specificity accurately, due to their different recognition methods for both target and PAM regions (Kleinstiver, *et al*., 2016).

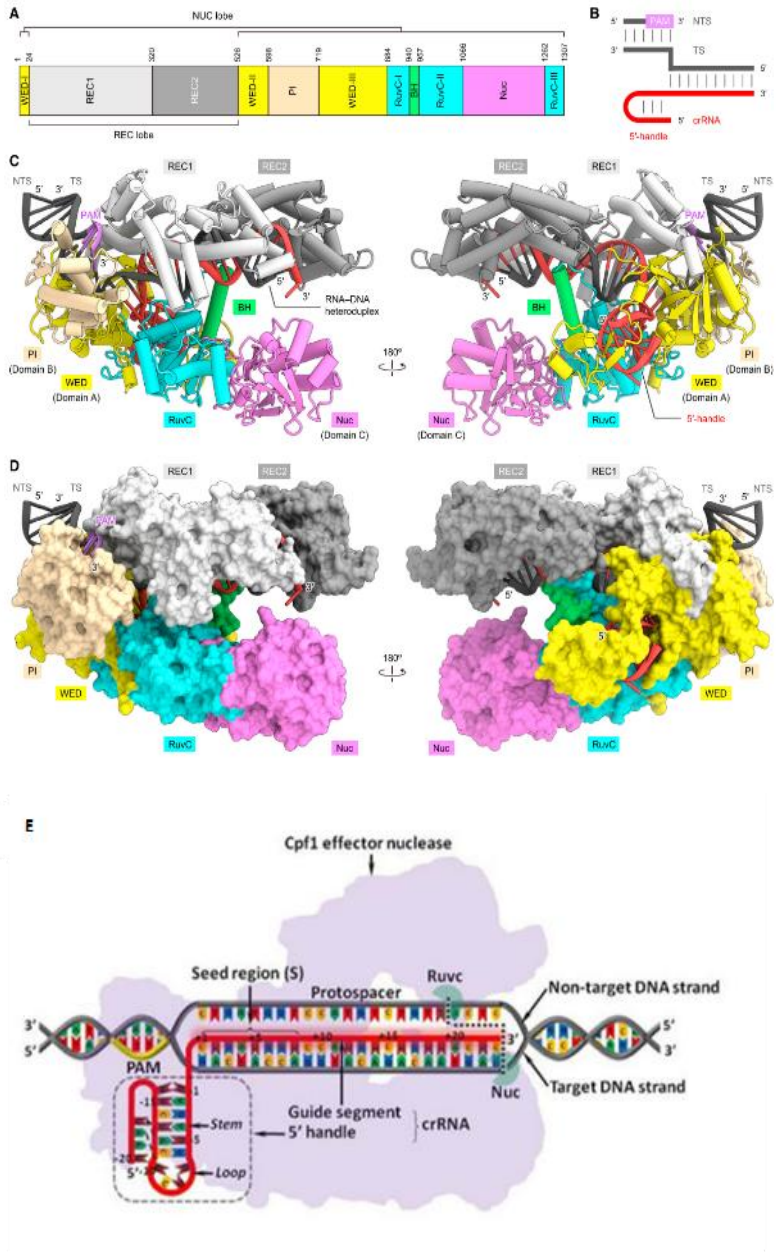


Figure 5 - Representation of the AsCpf-crRNA-DNA complex (AsCpf1 stands for Cpf1 of *Acidaminococcus* sp.). A - AsCpf1 domains' organization (with indication of NUC and REC lobes). B - crRNA (red) and target DNA (black) complex. The PAM sequence is represented in pink. C - Ribbon diagram of AsCpf1-crRNA-DNA ternary complex. D - Structure of the previously mentioned complex. The colors attributed follow the ones attributed in figure A (Yamano, *et al*., 2016). E - Representation of the AsCpf1-crRNA-DNA complex where the dotted lines indicate the complex cleavage sites (Li, *et al*., 2017).

Cas12a is composed of two main regions: REC (an alpha-helical recognition site) and NUC (a lobe with nuclease function) (Nishimasu, *et al*., 2015; Yamano, *et al*., 2016). The REC lobe consists of two domains: REC1 and REC2. The first is composed by 13 alfa-helices and the second by 10 alfa-helices and 2 beta-strands (Yamano, *et al*., 2016). According to Stella and colleagues, this lobe contains the regions with the higher level of

conformational flexibility (Stella, *et al*., 2018). The NUC lobe includes the domains WED (divided in 3 motifs), PI and RuvC (also divided in 3 motifs) (Figure 5A) (Yamano, *et al*., 2016; Stella, *et al*., 2018) and a bridge-helice, between RuvC-I and RuvC-II that connects both lobes (Nishimasu, *et al*., 2015; Yamano, *et al*., 2016). The PI domain is responsible for the identification of the PAM, its binding, and to promote the separation of the double strands of the target DNA (Stella, *et al*., 2018). It was observed that during the transition from the RNA-DNA-bound binary to ternary state, although Cas12a changes conformation, the RuvC and NUC domains (which constitute the active site of the protein) don't change their alignment (Gao, *et al*., 2016).

Cas12a crRNA has 42 to 44 nucleotides that can be divided in two sections: a variable guide segment (with approximately 23 nucleotides), complementary to the target DNA, and a 5'-handle (containing about 20 nucleotides), which is a conserved sequence arranged as a pseudoknot (Figure 5B and 5E) (Yamano, *et al*., 2016; Li, *et al*., 2017; Stella, *et al*., 2018). The latter can be identified by the WED domain of Cas12a, which is able to interact with the phosphate group between the last PAM nucleotide and the first nucleotide of the protospacer (Swarts & Jinek, 2018) - and bound between this domain and RuvC (Yamano, *et al*., 2016).

Kleinstiver and colleagues suggested that Cas12a might promote less off-target mutations that Cas9 nuclease (Kleinstiver, *et al.*, 2016). Nevertheless, some nucleases were already edited to improve their PAM specificity, such as AsCpf1 (Gao, *et al*., 2017; Nishimasu, *et al*., 2017). In the referred work, AsCpf1 variants were mutated to recognize TYCV and TATV PAMs (where "V" stands for adenine, cytosine or guanine and "Y" can be either cytosine or thymine), while retaining high specificity for their DNA targets (Gao, *et al*., 2017).

Some studies have already demonstrated that Cas12a, albeit being of prokaryotic origin, can be applied to the genome edition of several eukaryotic species. This endonuclease can be used to edit mammalians, such as humans and mice (Zhang, *et al*., 2017), yeast, insects, non-mammalian vertebrates (Swarts & Jinek, 2017) and plants (Wang, *et al*., 2018), revealing itself as a promising tool.

In a world where the human community is becoming more and more aware of the science done in research centres, the transmission of the achieved developments to the general population is becoming increasingly relevant. At the same time, ongoing and planned projects are being scrutinized, to confirm whether or not they follow the established ethical regulations. The edition of biomolecules (specifically proteins and

DNA) has been a controversial topic for many years and there are many who question where to draw the line between science and fiction. An example of such occurred on November of 2018, when the work of He Jiankui - who claimed to have manipulated the embryos of twin girls born in said month (Cyranoski & Ledford, 2018) - rose pertinent points, one of which was precisely the security of human embryos gene editing procedures, when considering the challenges previously mentioned - from which can be emphasised the off-target modifications (Chandrasegaran & Carroll, 2016).

With this work, the understanding of the catalytic mechanism of Cas12a will potentially improve, as will (hopefully) the enzyme specificity - what may be a further step to reduce CRISPR-Cas off target modifications, still a major drawback for this technology.

This work seeks to contribute for a more efficient and safe gene editing tool. With this, we hope to help the arrival of a time where genome editing is seen as a viable and sound alternative to drugs - especially in cases where said drugs only target the symptoms of a problem and not its roots - instead of a distant and unrealistic hypothesis.

Computational Methods

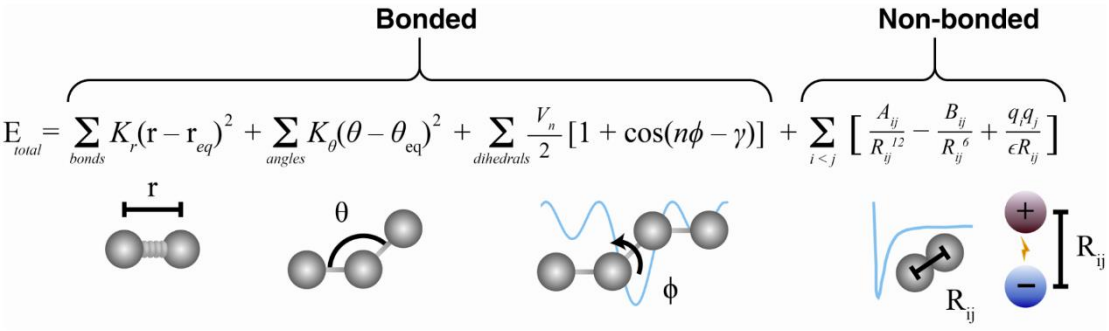
This study intends to achieve more atomic level insight on Cas12a's mechanism, to improve its specificity for a target strand. To achieve this objective, Molecular modelling, Molecular Dynamics (MD) simulations (Friesner & Gualar, 2005) and Quantum Mechanics/Molecular Mechanics calculations were performed.

In this chapter, the techniques used in computational chemistry will be briefly described.

Molecular Mechanics

Molecular mechanics is the name given to the methodology used to describe a molecular system through the application of simplified potential energy functions using empirical parameters. The term "force field" is given to the equation of the potential energy function and to the set of parameters applied to it (Durrant, *et al*., 2011).

The computational study of proteins considers several parameters, to keep the simulations as reliable as possible and try to accurately reproduce a molecule behaviour. These parameters intend to describe every particle, force and interaction that could be applied to a system. Said forces and interactions include the ideal length between atoms as well as their atomic angles, torsions and charges (Durrant, *et al*., 2011). The quality of results of a MD simulation, highly depends on the force field used and its capacity to accurately describe every factor and parameter applied to a system (Dourado, 2010). In the present work the force fields used were ff14SB and bsc1. The ff14SB force field is an evolution of ff99SB - it allows a better characterization of the backbone and side chain torsions for some amino acids than the previous model - and enables the characterization of proteins, nucleic acids and water molecules (Maier, *et al.*, 2015). The bsc1 force field is used to describe nucleic acids. Released in 2015, as an evolution of bsc0 (that in its turn was based on parm99, improving the α and γ dihedrals), it includes "additional modifications to the sugar pucker, the χ glycosidic torsion, and the ε and ζ dihedrals" and was developed to improve the performance of long MD simulations of double stranded DNA (Galindo-Murillo, *et al.*, 2016).



Equation 1 - Equation used for the AMBER force field (Durrant, *et al*., 2011).

The Equation 1 describes the parameters and forces applied to a molecule during a MD simulation, and is presented as the total energy of the system, this corresponds to energy variation promoted by changes in the connection between two atoms. The term $\sum_{bonds} K_r\ (r - r_{eq})^2$ describes the bonds between the system atoms with a simple harmonic function, where K_r is a force constant, r describes the distance between atoms and r_{eq} is an equilibrium structural parameter; $\sum_{angles} K_\theta(\theta - \theta_{eq})^2$ describes the angles established between bonds, similarly to what happens for the first term, this parameter is also described by a harmonic function and K_θ corresponds to a force constant, where the angle between the bonds is represented by θ and θ_{eq} is another equilibrium structural parameter; $\sum_{dihedrals} \frac{V_n}{2}[1 + \cos(n\emptyset - \gamma)]$ refers to the dihedrals between angles and bonds, represented with a truncated Fourier expansion, where V_n is a force constant and n is the multiplicity of an energy level and γ is "phase angle for torsional angle parameters" (Wang, 2006), while the parameter $\sum_{i<j}[\frac{A_{ij}}{R_{ij}^{12}} - \frac{A_{ij}}{R_{ij}^{6}} + \frac{q_i\, q_j}{\epsilon R_{ij}}]$ is used to describe the behaviour of non-bonded atoms, connected, for example, with van der Waal interactions, in accordance to Lennard-Jones Potential (Equation 2) and Coulomb's Law (Equation 3) explained below (Durrant, *et al*., 2011, Salomon-Ferrer, *et al*., 2013).

$$V(r) = 4\epsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right]$$

Equation 2 - Lennard-Jones Potential (Naeem, 2019).

The Lennard-Jones Potential describes the potential energy of interaction - or intermolecular potential - between two structures (such as atoms or molecules) with base on the distances between them. The two sections of the second term of equation 3 ($\left(\frac{\sigma}{r}\right)^{12}$

and $\left(\frac{\sigma}{r}\right)^6$) refer, respectively, to the repulsive and attractive forces between particles (Naeem, 2019).

$$F = k\frac{q_1 q_1}{r^2}$$

Equation 3 - Coulomb's Law (Haas, 2019).

Electrostatic interactions are calculated using Coulomb's Law. It is applied to describe attractive and repulsive forces between electrically charged particles (Haas, 2019). It demonstrates that "two like charges repel each other with a force that varies inversely as the square of the distance between them" (Bartlett, *et al.*, 1970).

Quantum Mechanics

Quantum mechanics is a methodology used to describe the behaviour of microscopic "particles" within a system. It defends that by knowing a set of characteristics of a "particle" at a certain time in a classical-mechanical system, it is possible to predict its future state, even if not every variable may be accounted for, as presented by Heisenberg in his uncertainty principle (which says that it is not possible to predict both the velocity and the position of a microscopic particle at the same time) (Levine, 2013)

With this in mind Schrodinger developed an equation that attains to describe the movement of a single particle within an one dimensional system - by predicting the probabilities of said particle being in a determined place when following a wave like motion described by a wave function.

$$-\frac{\hbar}{i}\frac{\partial\Psi(x,t)}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2\Psi(x,t)}{\partial x^2} + V(x,t)\Psi(x,t)$$

Equation 4 - Equation of Schrodinger (Levine, 2013).

In Equation 4, Ψ is considered the function of the particle coordinates (and has in account both time and space), i equals to $\sqrt{-1}$, m represents the mass of the particle, V(x,t) is the potential energy function of the system and $\hbar$ is defined by $\hbar \equiv \frac{h}{2\pi}$, where h is Planck's constant (Levine, 2013).

However, this equation may only be applied to a single particle and it was necessary to develop more advanced formulations in order to characterize the whole system. These

include the Born-Oppenheimer approximation, which makes possible the separation of nucleic and electronic motions (Kosloff, 1988), the Hartree-Fock formalism, among others. In this project it was used the semi-empirical model PM6, a more refined version of the systems AM1 and PM3 (Christensen, *et al*., 2016).

Quantum Mechanics/Molecular Mechanics

QM/MM is a hybrid method used to study large molecules. This combination of both methodologies allies precision with velocity, allowing the treatment of (in the case of an enzyme) the catalytic region of a molecule with QM and the remaining residues, as well as the solvent solution with MM (Levine, 2013).

$$\hat{H}_{\text{eff}} = \hat{H}_{\text{QM}} + \hat{H}_{\text{QM/MM}} + \hat{H}_{\text{MM}}$$

Equation 5 - Hamiltonian description of a QM/MM system.

In the Equation 5, $\hat{H}_{QM}$ is the Hamiltonian which describes the part of the system characterized with QM, $\hat{H}_{MM}$ describes the section of the system described with MM and $\hat{H}_{QM/MM}$ represents the interaction between both regions.

Molecular Dynamics

There are several methodologies that enable the study of a molecule' behaviour, movement and mechanics. In this work, the strategy used is Molecular Dynamics (MD). In MD, at each time step, new parameters are calculated for each particle, using as reference the values obtained for the previous step, in accordance to Newton laws of motion.

$$\frac{1}{2}\sum \boldsymbol{m}_i v_i^2 = \frac{3}{2}\boldsymbol{N}\boldsymbol{k}_{\mathrm{b}}\boldsymbol{T}$$

Equation 6 - Representation of the Classical Model, used to develop MD simulations (Levitt, *et al*., 1995).

Equation 6 represents a determined system in an equilibrium situation. In this, it is possible to see that the velocity (v) of a particle (i) is related with the temperature (T) of the system through its' total kinetic energy ($\frac{1}{2}\sum m_i v_i^2$). m_i represents the mass of each

particle, while k_b is the Boltzman constant and N the total number of atoms in the system (Levitt, *et al.*, 1995).

Enzymes' catalysis often occur in an aqueous environment, hence the importance of the construction of reliable parameters and models to describe this environment - that should be considered when running computational simulations (Jorgesen, *et al*., 1983; Friesner & Gualar, 2005). The protein dynamics performed used the solvation model TIP3P within an octahedral box (Almeida, *et al*., 2019; Jorgesen, *et al.*, 1983), in explicit solvent - water (Salomon-Ferrer, *et al*., 2013) - with periodic boundary conditions.

Periodic boundary conditions allow the system to be at a stable pressure while its volume varies during a MD simulation - what will allow an equilibration of every component of the same (Madej, *et al*., 2015; 2018). Using this strategy, one may consider that the structure in study is enclosed in a box of the chosen solvent, however, in order to prevent the interaction of the solvent with a hypothetical atmospheric layer on the outside of the "box" - what would disrupt the study, since said interaction would have to be considered - periodic boundaries conditions are used. These consider that the "box", which contains the solvent and the molecule of interest, is surrounded by similarly built "boxes", in a three dimensional context (as is represented in two dimensions in Figure 6). This ensemble contemplates that the behaviour of each particle is exactly the same at the same time point for every solvent box, what would mean that every time a particle A goes beyond the boundaries of a box α, a particle B would reflect that movement and enter the box α, leaving a box β, where the same would happen (Madej, *et al*., 2015; 2018).

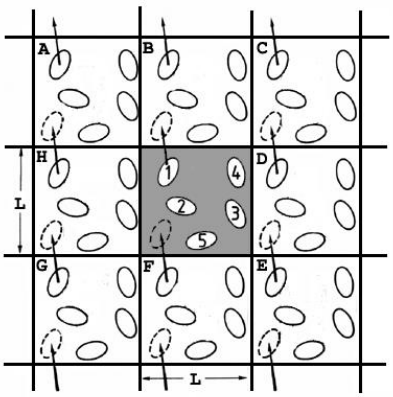


Figure 6 - Scheme of a molecule periodic boundary conditions (Madej, *et al*., 2015; 2018).

In the present work, the cut-off used was of 10 Å. This value indicates the distance at with non-bonded interactions should stop being calculated and needs to be wide enough to include the van der Waal interactions, while small enough to reduce significantly the

computational cost of a simulation. This is necessary to specify due to Particle Mesh Ewald (PME) method (Case, *et al*., 2018). While this method introduces periodicity to the system, what may be more stable than vacuum boundaries (Souza and Ornstein, 1997) and "an efficient way to compute long-range forces" (Fuxreiter, 2014), at the same time it increases the computational costs, as it also generates "infinite electrostatic" interactions (Madej, *et al*., 2015; 2018).

In Molecular Dynamics simulations, the structure used as reference for the performance of a simulation is subjected to several procedures before the start of the dynamics. Firstly, the files that will be used have to be prepared and solvated - in the present work this was done with the AMBER programm LEAP. The second of the referred procedures is called minimization. This step optimizes the system geometry by minimizing its energy, allowing its particle to reach its lowest energy value. During this stage, the enzyme and the nucleic acids are subjected to restraints and its volume is kept constant. After minimization, the system is subjected to a "heating", or "equilibrium" stage, where its temperature is risen from 0 to 310,15 kelvin degrees - bringing the system to a temperature compatible with that of biological systems (in our case, 37 Celsium degrees (ºC)) (Case, *et al*., 2018). In this stage, the velocity of the particles of the system relates with the temperature, increasing with it - as may be observed in equation 1 (Levitt, *et al*., 1995) - and the environment of the protein is kept at a constant pressure, while its volume may vary. Both the second as well as the third stages were performed using the AMBER programm *sander*. To the fourth stage is called production. This is the phase where the commands are given to start the simulation in itself, this is represented in Figure 7.

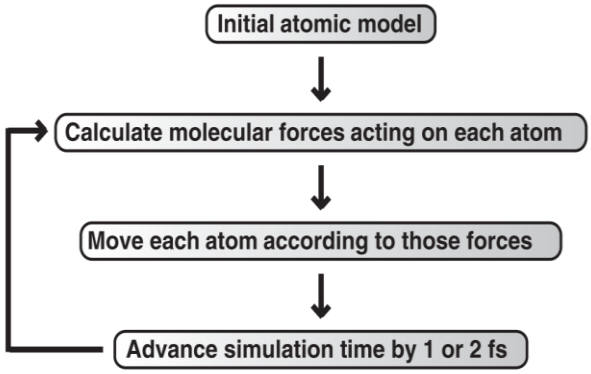


Figure 7 - Schematic representation of a MD simulation (Durrant, *et al*., 2011).

Analysis of a MD trajectory

There are several methods to analyse a MD trajectory. Here will be presented the most relevant for this work.

RMSD and RMSF were used to perform preliminary analysis of the system, to compare the mutant enzymes' stability with the same parameters for the wild type molecule.

RMSD

The root mean squared deviation value compares how similar a structure's internal atomic coordinates are in relation to a reference molecule coordinates (Case, *et al*., 2018). It has several applications, including the evaluation of structural changes in protein dynamics (Kuzmanik & Zagrovic, 2010). RMSD is represented by the following equation (Case, *et al*., 2018):

$$RMSD = \sqrt{\frac{\sum_{i=0}^{N}[m_i * (X_i - Y_i)^2]}{M}}$$

Equation 7 - Root Mean Square Deviation equation.

In this equation, N represents the number of atoms of the system, m_i stands for the mass of each atom (i), X_i and Y_i are the coordinate vectors for the target and reference atoms, respectively, and M the total mass of the system (Case, *et al*., 2018).

RMSF

The root mean square fluctuation, also called "atomic positional fluctuations" captures the fluctuations of each residue of the structure during the simulation relatively to a reference position, being a fast way to assess the flexibility of the enzyme as well as its stability (Arodola & Soliman, 2016; Gromacs Tutorial, n.d.).

$$\mathrm{RMSF}_i = \left[\frac{1}{T}\sum_{t_j=1}^{T}|\mathbf{r}_i(t_j) - \mathbf{r}_i^{\mathrm{ref}}|^2\right]^{1/2}$$

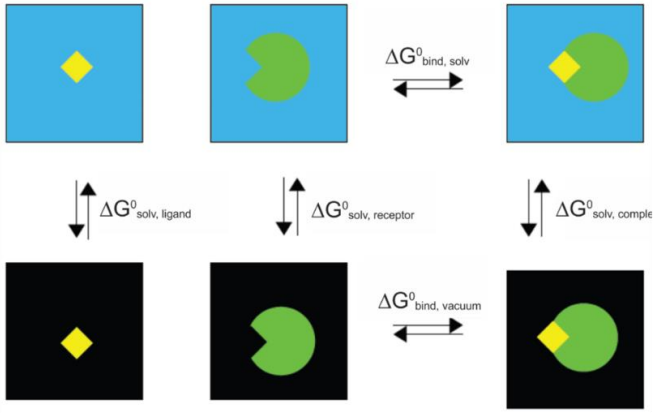
Equation 8 - Root Mean Square Fluctuation equation.

In this equation, T represents the time period selected for the analysis and r^{ref} is the reference position of a particle i (Gromacs Tutorial, n.d.).

MMPBSA

MMPBSA (Molecular Mechanics - Poisson Boltzmann Surface Area) is one of the methods used to calculate the binding free energies between a receptor and a ligand. It was decided to use this method in the present work due to its simplicity and relatively fast calculations. The free energies calculation is a useful tool for drug design, protein structure determination and to determine the stability of several conformations of a protein (here it was used to determine the stability of the system), however this process is also computationally expensive. MMPBSA allows the reduction of this cost by using intermediate structures - snapshots of previously performed MD simulations, instead of running them solely for this purpose - and by reducing the disturbance caused by explicit solvent molecules - by comparing structures in vacuum with their homologous states in a dynamic where the explicit solvent molecules are present (Miller, *et al.*, 2012; Poongavanam, *et al.*, 2014) - otherwise the amount of solvent molecules present in solution and the energy contributions from their interactions would overwhelm the binding energy between the receptor and the ligand, drastically reducing the comprehension of the results (AMBER Advanced Tutorials - Tutorial 3).

Binding free energies are calculated by comparing the unbounded energy of both the receptor (in our case, the protein Cas12a) and the ligand (the nucleic acids which complex with Cas12a - DNA and RNA) with the free energy of the bounded complex in vacuum and using and implicit solvent model, as represented in the thermodynamic cycle in Figure 8.



$$\Delta G^0_{bind,\,solv} = \Delta G^0_{bind,\,vacuum} + \Delta G^0_{solv,\,complex} - (\Delta G^0_{solv,\,ligand} + \Delta G^0_{solv,\,receptor})$$

Figure 8 - Schematic representation of the thermodynamic cycle for binding free energies calculations. $\Delta G^0_{bind,solv}$ and $\Delta G^0_{bind,vacuum}$ correspond to the free energy differences between the bounded and the unbounded states of the complex solvated and in vacuum and $\Delta G^0_{solv,ligand}$, $\Delta G^0_{solv,receptor}$ and $\Delta G^0_{solv,complex}$ indicate the free energy changes of the solvated and vacuum states of the referred structure The square in yellow represent the ligand and the shape in green the receptor. The boxes in blue represent solvated systems and the boxes in black represent in vacuum systems.

Wet Lab Methods

This phase of this work aimed to validate the most efficient of the Cas12a variants generated during the computational stage in comparison with the WT enzyme used to develop the referred mutants and considered the control test.

In this stage I followed the "Heterologous Expression and Purification of CRISPR-Cas12a/Cpf1" protocol (Mohanraju, *et al*., 2018). It consists in three main tasks: 1 - Expression of the protein in the *Escherichia coli* (*E. coli*) strain BL21 Star (DE3) (initially it will be expressed the WT enzyme and once the protocol is optimized, we will proceed with the expression of the most promising Cas12a mutant developed during the computational stage of this work); 2 - enzyme purification (similarly to the previous task, first protocol will be optimized for the WT enzyme and afterwards it will be applied to the expressed mutant); 3 - Activity assay using the purified enzymes.

Due to SARS-CoV-2 restrictions, the experimental phase of my work was compromised and it was only possible to advance with the first main task of the chosen protocol, and even so, it was solely applied to the plasmid containing the WT enzyme, in order to assess the functionality of the selected protocol and to allow for optimizations and adaptations of the referred protocol to the laboratorial conditions and available material.

The sequence of the protein of interest was received in bacterial swabs from Addgene (pMBP-FnCas12a - ID 113432) and did not require further alterations before the transformation. The first task was encompassed in six steps: 1 - Growth of bacterial cultures, 2 - Preparation of plasmid Miniprep solutions to amplify plasmid DNA for storage, 3 - transformation of *E. coli* BL21 Star (DE3) with the expression plasmid pMBP-FnCas12a, 4 - Preparation of glycerol cell stocks; 5 - plasmid DNA confirmation by DNA sequencing and 6 - expression of MBP-FnCas12a fusion protein in *E. coli* BL21 Star (DE3) cells, followed by polyacrylamide gel electrophoresis (SDS-PAGE) analysis of the expression product.

Growth of bacterial cultures

The plasmid pMBP-FnCas12a has 9869 base pairs. It contains a T7 promoter, a 10-His tag, a sequence encoding for the fusion protein Maltose Binding Protein (MBP) from *E.coli*, a Tobacco Etch Virus (TEV) protease site (which will allow the removal of the MBP tag), the protein of interest for this work - FnCas12a - and a T7 terminator. This protein has 1300 amino acids, corresponding to 3900 base pairs, and a molecular weight of 155 kDa - 197,3 kDa if it still contains its affinity tag, MBP, and the 10-His tag, before cleavage with TEV protease . This plasmid habors a gene which encodes for a protein which confers resistance to ampicillin to the organism containing it.

This plasmid was received in the laboratory as bacterial swabs and in addition to the information that our plasmid of interest was resistant to ampicillin, the vial it was received in also indicated the plasmid was inserted in the growth strain *E. coli* DH5alpha. After harvesting, the cells were grown in both solid LB agar and liquid LB mediums. Three replica were made on petri dishes, two in LB medium containing ampicillin and another in LB medium containing kanamycin, acting as a negative control, another three replica were made by adding the harvested cells to 10 mL of liquid LB medium on three different falcon tubes. In parallel and similar approach was done with solid agar medium: two LB-Agar plates supplemented with ampicillin and one LB-Agar plate supplemented with kanamycin (control) were used.

Preparation of plasmid DNA stocks (DNA Minipreps)

After the previous step, the colonies containing the plasmid of interest - resistant to ampicillin - were selected to inoculate new cultures (through single colony picking) in order to obtain saturated cells cultures for glycerol stock preparation for long term storage at -80ºC.

Plasmid DNA stocks were prepared as described below.

- Plasmid DNA stocks (pDNA) - For this solution was followed the NZYMiniprep kit protocol (product reference ID: MB010) for plasmid DNA purification from *E. coli* cells. 10 mL of the liquid cultures grown from the bacterial swabs were transferred to 15 mL falcon tubes and centrifuged for 10 minutes at 4200 rpm, being the supernatant discarded. Pellet was re-suspended with 250 µL of Buffer A1 using a micropipette after which was added 250 µL of Buffer A2 and gently mixed, followed by the addition of 300 µL of

Buffer A3, which was also carefully mixed. The next step was the centrifugation of the sample for 10 minutes, also at 4200 rpm. Supernatant was load to NZYTech spin columns - inserted in 2 mL collecting tubes - centrifuged at 11000 rpm for 1 minute and the flow-through discarded. After this, 600 µL of Buffer A4 were added to the column and centrifuged under the same conditions as the previous step, the flow-through also was discarded. Next step consisted on transferring the NZYTech spin columns to empty 2 mL collecting tubes and their centrifugation for 2 minutes, followed by the placement of the dried column in clean 1.5 mL microcentrifuge tubes to which were added 35 µL of autoclaved water, these were left to incubate for a minute, after which were centrifuged for another minute. Plasmid DNA concentration was measured using a Nanodrop ND-1000 Spectrophotometer and stored at -20ºC.

Heat-shock bacterial transformation

In this step, the pDNA obtained in the previous stage was inserted into competent *E. coli* BL21 Star (DE3) bacterial strain. For this to occur, a 100 µL aliquot of inhouse prepared competent cells was retrieved from a -80 ºC chamber and left to thaw for 5 minutes on ice. Once the cells reached the intended state, 1 µL of pDNA was added and this mix was left to rest for 30 minutes on ice, after which the cells were subjected to heat shock treatment for 1 minute, using a thermoblock at 42 ºC. Cells were left to recover for 30 minutes on a shaker at 37 ºC and 160 rpm in 1 mL of LB medium and subsequently centrifuged for 3 minutes at 2600 rpm. Cells were then re-suspended in approximately 100 µL of supernatant, and plated on LB agar plates containing ampicillin and incubated at 37 ºC during 16 hours. On the next day, colony-containing plates were transferred to a 4 ºC chamber and kept until further use.

Preparation of glycerol cell stocks

Glycerol stocks of transformed *E. coli* BL21 Star (DE3) with the plasmid pMBP-FnCas12a were prepared as described below.

The cells grown in the petri dish were re-suspended using 5 mL of LB medium. After this step, 800 µL of the medium with the re-suspended cells were transferred to another vial were was added 200 µL of glycerol, resulting a final concentration of 20%. This stock solution was flash frozen and kept at -80 ºC.

Plasmid DNA Sequencing

In order to double check if the plasmid DNAs (pMBP-FnCas12a - ID 113432) received from Addgene were the requested ones, two samples freshly inhouse prepared DNA plasmids were sent for sequencing at GATC Services - Eurofins Genomics (Germany). Premixed samples containing 5 µL of plasmid DNA with 5 µL of sequencing oligonucleotides were prepared according to GATC recommendations. Standard sequencing oligonucleotides working from T7 Promoter and T7 Terminator regions were used to sequence respectively through the 5' end and 3' end of the MBP-FnCas12a encoding sequence.

Heterologous expression of MBP-FnCas12a recombinant fusion protein

For protein expression, the transformed cells were retrieved from glycerol stocks. Said cells were collected with a micropipette tip and used to inoculate 10 mL of LB medium containing 10 µL of ampicillin. This pre-culture was incubated in a shaker at 37 ºC and 160 rpm during 16 hours, after which, 9 mL of the culture were added to a 5 L Erlenmeyer containing 1 L of LB medium and 1 mL of ampicillin (100 mg/mL), this culture was incubated on a shaker at 37ºC and 160 rpm until reaching an OD of 0.5 - 0.6. When this OD values were reached, the culture was subjected to a cold-shock for 15 minutes on ice and 133 µL of IPTG (Isopropyl β-D-1-thiogalactopyranoside) (1M solution) were added to induce protein expression. The cultures were transferred to a shaking incubator for approximately 16 hours at 120 rpm and 18 ºC - as this protein is relatively big, a lower temperature helps its correct folding.

Analysis of MBP-FnCas12a expression by SDS-PAGE

For this, a sample of 1 mL of the culture was collected before expression induction with IPTG (T0 time point) and another sample was collected after the 16 hours of growth (T_{ON} time point). The Optical Density (OD) of each sample was measured and the sample centrifuged at 1300 rpm for 9.5 minutes, being the supernatant discarded and the pellet re-suspended in Phosphate-buffered saline (PBS) - the OD was used to normalize the amount of cells, and consequentially of protein, per sample. This step allows a correct comparison between T0 and T_{ON} values, emphasizing the overexpression of the protein.

In a new eppendorf tube, 15 µL of loading buffer was added to 30 µL of the re-suspended bacterial pellet with PBS - this samples was kept at 4 ºC until further use.

To run an electrophoresis in polyacrylamide gel it is necessary to prepare said gel. It was decided to prepare one consisting in two sections: a stacking gel, at 6% of acrylamide, and a running gel, at 12% of acrylamide.

Before loading onto the polyacrylamide gel, lysate samples were boiled at 95 ºC to ensure effective protein denaturation. Once it was guaranteed, the gel was charged with 6 µL of a Precision Plus Protein™ Unstained Protein Standards marker (Biorad) and with 15 µL of each sample per well.

The electrophoresis run started with a voltage of 120 V, until the samples reached the running gel, after which it was increased to 160 V, for 80 minutes.

Computational Results and Discussion

Molecular Modelling

The initial structure for the simulation was retrieved from the RCSB Protein Data Bank archive. This data base keeps information about Biological Macromolecular Structures, which include the sequences and three dimensional shapes of proteins and nucleic acids (RCSB, n.d.).

This structure pdb code was 6GTF (RCSB - 6GTF, n.d.). Deposited on the 18th of june of 2018 by Montoya G. and colleagues (Stella et al., 2018), this conformation of Cas12a was obtained through electronic microscopy with a resolution of 3.63 Å. The downloaded file contains information on 3 sequences: the aminoacid sequence of the protein - 1329 residues -, the CRISPR RNA (crRNA) - 43 residues - and the target DNA sequence, which contains both the template and the non-template strands - 55 residues (Berman, *et al*., 2000). 6GTF structure (called I5 conformation in Stella and colleagues article) was chosen for being the structure immediately before the chemical step (Stella, *et al*., 2018).

The retrieved structure lacked Mg^{2+} ions, had incomplete DNA strands (the non-template chain did not reach the cleavage site), the glutamate 1006 (which was present in 6GTF as a glycine, otherwise the enzyme would be in its active state - with GLY1006, Cas12a is an apo-enzyme - and it would not be correctly represented in the crystal) and it did not include hydrogen atoms. This meant that the structure needed to be modelled and corrected before it was possible to further advance with the project.

The modelling and alignment of the DNA strands was performed with Chimera, where the nucleotides missing were added one by one, guaranteeing the complementarity of the template and non-template strands. It was also Chimera the program used to exchange the GLY1006 present in the 6GTF structure with the glutamate, which was absent. The missing hydrogen atoms were added in accordance to the residues protonated designations. In this project, the magnesium ions acting as cofactors for the reaction (Zetsche, *et al*., 2015) where placed in the structure considering their hexacoordinated geometry. Considering also the similarity of domains between FnCas12a and the RuvC domain of Cas9, this work was developed with the assumption that the catalytic activity of our enzyme occurs through a two-metal ion mechanism (Palermo, 2019). This placement was done using the Amber parameters for non-bonded ions and the knowledge

that these ions are crucial for the catalysis of the enzyme, as they coordinate the attack of the nucleophilic water and stabilize the leaving group (Yang, *et al*., 2006).

After these corrections, the pdb file obtained was geometry optimized by the Amber18 program, with the ff14SB and bsc1 force fields (Figure 9).

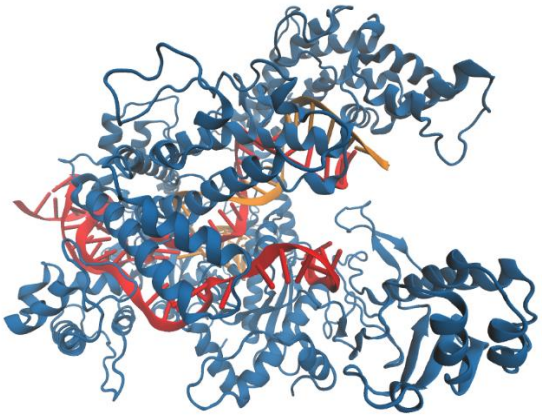


Figure 9 - Representation of Cas12a (blue) with its' target DNA (red) and CRISPR RNA (orange).

The mutants created during this project were obtained by substituting chosen amino acids of the referred structure by another residue, in order to either obtain structures with increased specificity for an intended target (N124K, N856K, V862K, R1218K, N124K_N856K_V862K) or to verify their relevance for the catalytic activity of the enzyme (H922A). After performing the intended substitution for each mutant created, the new structures geometry was optimized, with Amber18 and the force fields ff14SB and bsc1.

Specificity improvement

Mutations selection

As previously mentioned, this work was divided in two sections, one dedicated to the improvement of Cas12a specificity and the other to the understanding of the enzyme catalytic mechanism.

After modelling the structure of our protein, it was necessary to assess which residues to exchange to achieve our goals. Visual inspection of the optimized structure and of a short MD simulation allowed to select residues of the WT enzyme that could be

exchanged in order to achieve a higher number of interactions with the nucleic acids, what could lead to an increase of the enzyme specificity. According to Dong and colleagues, the recognition of the crRNA by the enzyme is important for target binding and the pseudoknot formed in the 5' terminal of this chain relevant for it to keep its conformation (Dong, *et al.*, 2016). With this in attention, it was considered that it would be interesting to observe the effects of a higher number of interactions with this site, and thus, residues from the crRNA loop were chosen as targets for protein residues to interact with. From Yamano *et al*., we learned that the last two nucleotides of the PAM motif are not recognized in a base-specific way, what allows a higher flexibility of recognition (Yamano *et al*., 2016). In addition, we were curious about the role of lysines for this motif recognition, and decided to select a neutral residue near the PAM to exchange by a lysine to study whether it would affect the specificity of the molecule.

A list of the residues located close to nucleotides of interest - PAM region and crRNA loop - was created and used to evaluate the distance between the chosen residues and the nucleic acids targets during a 5 nanoseconds simulation.

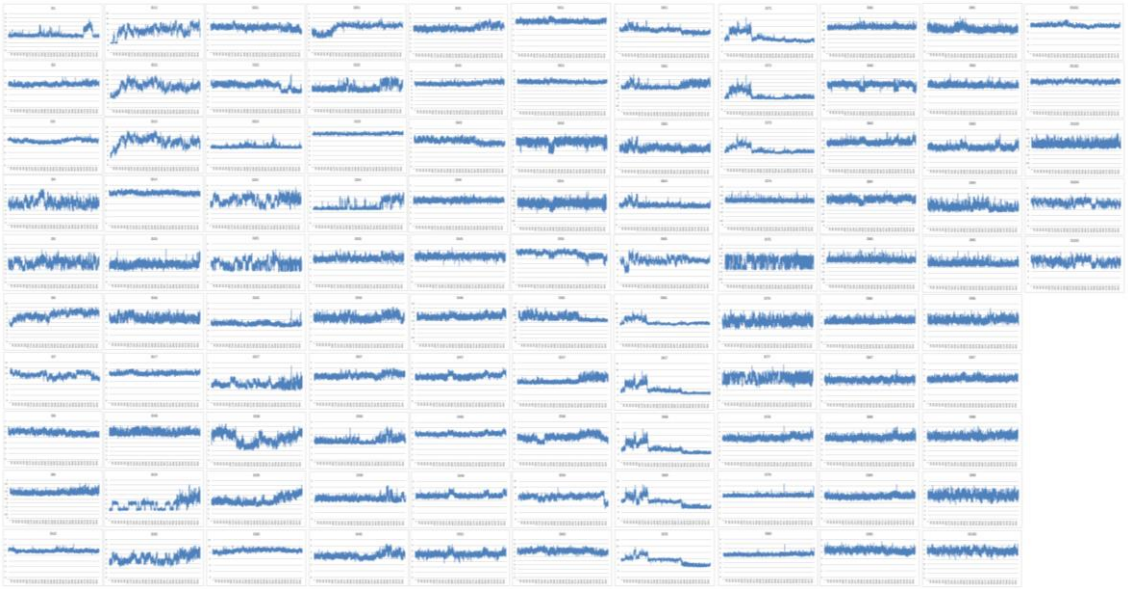


Figure 10 - Representation of each of the 105 distances analysed.

From the 105 distances analysed (Figure 10) for different atoms belonging to 26 amino acids and 8 nucleotides (for further information see Anex II - Part1), three protein residues were selected and subjected to in silico mutagenesis.

To create stronger interactions between the protein and the nucleic acids, charge-charge (or salt-bridge) interactions were established, by substituting a non-polar valine

and two asparagines by lysine residues. The residues chosen were the Asparagine 124, close to the PAM sequence on the non-template strand, the Asparagine 856, close to the crRNA loop, and the Valine 862, also close to the crRNA loop.

With the selection of these residues, it was decided to develop four mutants, one targeting each residue (N124K, N856 and V862K) and a fourth which would combine the three mutations (N124K_N856_V862K). The selected residues were exchanged to lysine residues. The choice of this amino acid was based on its positive charge (an advantage, when the mutation goal is to increase the protein interaction with nucleotides which are negatively charged at the backbone), long structure and high number of hydrogens at the end of its lateral chain - increasing the chances of creating hydrogen bonds.

The obtained results were divided in two sections: 1) mutants created to increase the number of interactions between Cas12a and the PAM region - presented in the section named “PAM” and 2) mutants developed to increase the number of interactions between Cas12a and the crRNA 5’ handle - presented in the section named “crRNA”.

PAM

The PAM motif is a short DNA sequence, which identifies foreign genetic material, once this motif is not present in the spacers acquired during the adaptation phase of the CRISPR-Cas system defense mechanism. Cas12a recognizes a thymine rich spacer, in the present case corresponding to the sequence TTA of the non-template DNA strand.

Two mutants were developed targeting the PAM region: N124K and the N124K_N856K_V862K. The goal of these mutations was to increase the number of interactions between our enzyme and the PAM motif of the target strand of DNA. In the N124K mutant, the distance to DT11 decreases to around 2.17 Å, since the hydrogen bond between the NH2 group of the N124 and the charged oxygen of the phosphate backbone of DT11 is substituted to a charge-charge interaction with the lysine, as may be seen in Figure 11.

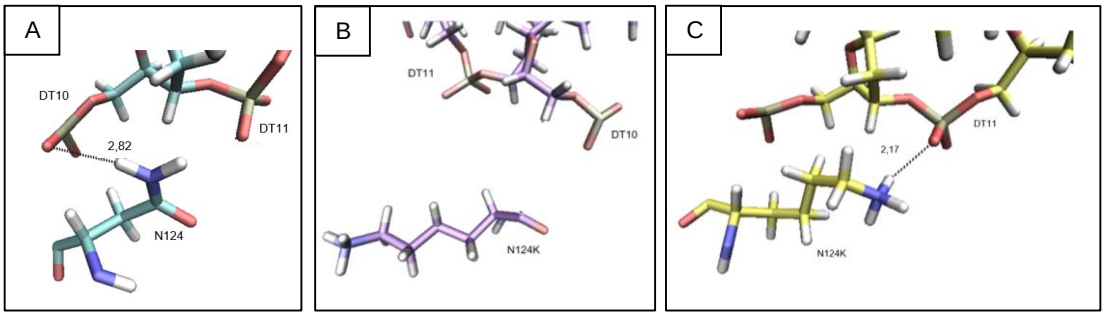


Figure 11 - Representation of the residue 124 of the (A) wild type, (B) N124K_N856K_V862K and (C) N124K structures and its closest DNA nucleotides, the values presented are in Ångström (Å).

To understand whether this mutation induced instability on the enzyme, RMSD and RMSF analysis were performed for the Cα residues of the protein, as is presented in Figure 12.

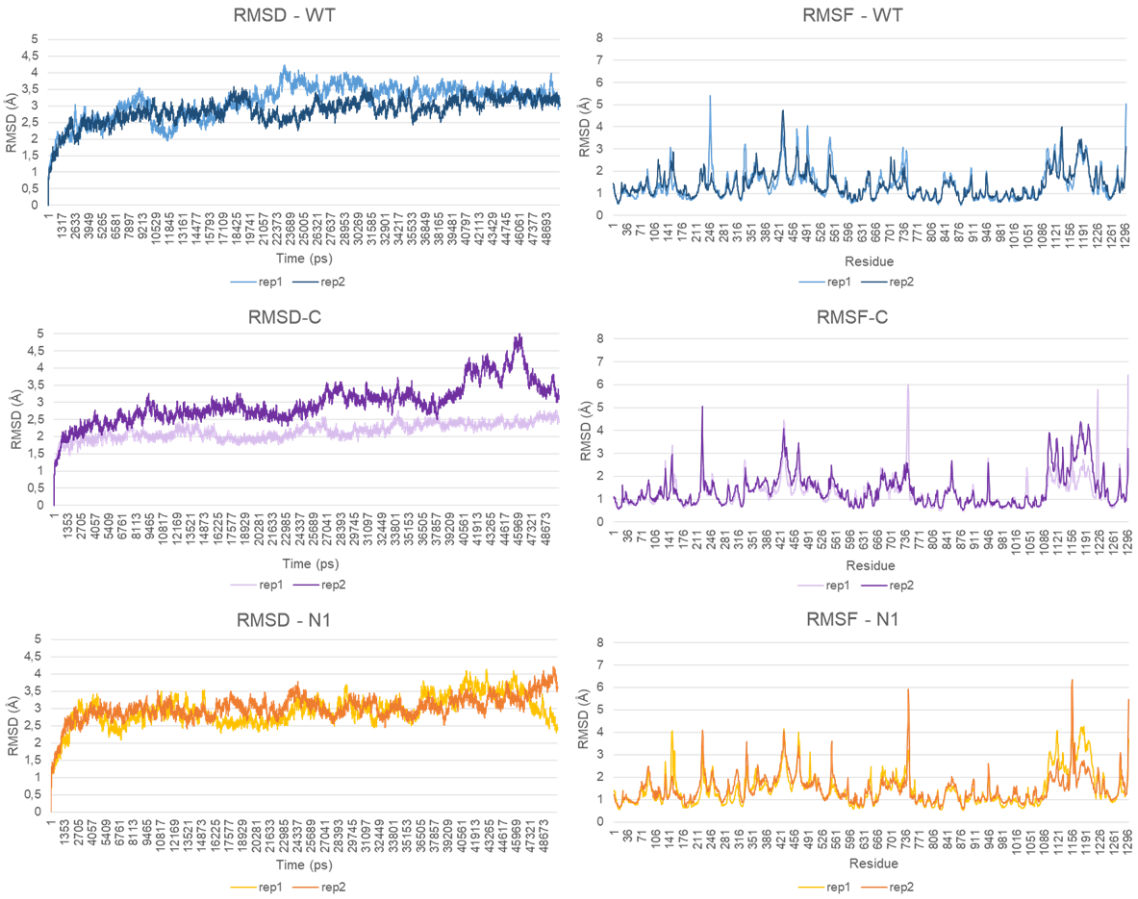


Figure 12 - Representation of Root Mean Square Deviations (RMSD) and Root Mean Square Fluctuations (RMSF) graphics for the first and second replica of the wild type structure (WT) (rep 1 and rep2, in light blue and dark blue, respectively), the N124K_N856K_V862K (C) (rep 1 and rep2, in lilac and purple, respectively, as well) and the N124K mutant (N1) (rep 1 and rep2, in yellow and orange, respectively).

From the obtained values for the RMSD and RMSF, it is possible to observe that the N124K_N856K_V862K mutant presents the highest value of RMSD Both the N124K mutant and the WT enzyme stabilize after approximately 3 nanoseconds of simulation.

From the RMSF graphs, it is noticeable that the structures present different spikes among them. In the RMSF analysis, for both N124K and N124K_N856K_V862K there is a spike at the residue GLN745, this residue makes part of one of the external loops of the enzyme, being part of the WED domain; there is also a pronounced spike for the N124K mutant at the residue SER1157, a loop belonging to the RuvC domain and one other for the N124K_N856K_V862K structure, at the GLY1238, at an external loop of the Nuc domain.

As previously mentioned, the goal of these mutations was to increase the number of interactions between Cas12a and the PAM site, what was done by creating charge-charge interactions. To compare this interactions to those which were present on the wild type structures, it was decided to evaluate the distances and angles between different atoms belonging either to the protein or to the nucleic acids (The selection done may be consulted in Anex II - Part2) (Figure 13). The atoms chosen were selected for analysis through observation of the final frame of each MD simulation). Here, we consider that there is a hydrogen bond interaction if two atoms (one of which a hydrogen connected to a heavy atom of a residue and a heavy atom from another residue) are closer than 3 Å apart and at the same time make an angle between 140º and 180º.

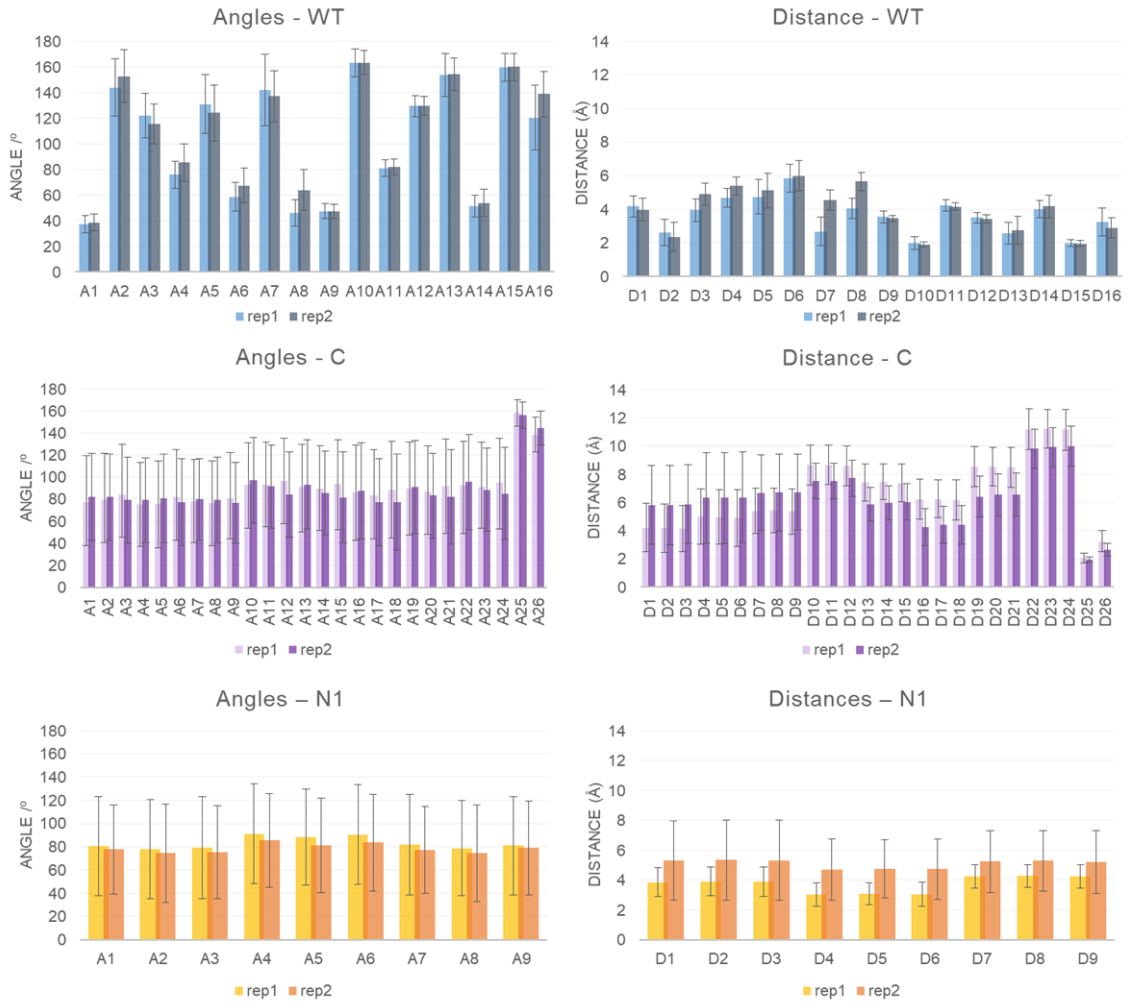


Figure 13 - Representation of mean values of the angles and distances, for the first and second replica of the wild type structure (rep1 and rep2, in light blue and dark blue, respectively), the N124K_N856K_V862K mutant (C) (rep1 and rep2, in lilac and purple, respectively, as well) and the N124K mutant (N1) (rep1 and rep2, in yellow and orange, respectively).

As it is possible to distinguish when consulting the second section of Anex II - Part2, the angle and distance values analysed for N124K correspond to those from 1 to 8 of the WT structure and to those from 1 to 9 of the N124K_N856K_V862K mutant, for both the distances and the angles. From the values obtained, we can see that the WT structure is the one who presents the more variation, with some atoms establishing interactions with the nucleic acids - such as the hydrogen bond between N124 atom HD22 the nucleotide DT10 atom OP1 (A2 and D2) and the N124 atom HD21 the atom OP2 of the nucleotide DT11 (A7 and D7) while others do not. It is possible to observe that for the mutants with the charge-charge interaction, the mean values of both angles and distance values is more uniform than in the wild type structure.

crRNA

Three mutants were developed targeting the 5'-handle of the CRISPR-RNA, N856K, V862K and the N124K_N856K_V862K, which is also mutated in both residues. These mutants were created with the intention to increase the number of interactions between Cas12a and this loop of the crRNA. From the Figure 14, it is possible to see that the charge-charge interaction created is stronger that the hydrogen bonds previously present, what brings the residue 856 closer to the U10 nucleotide.

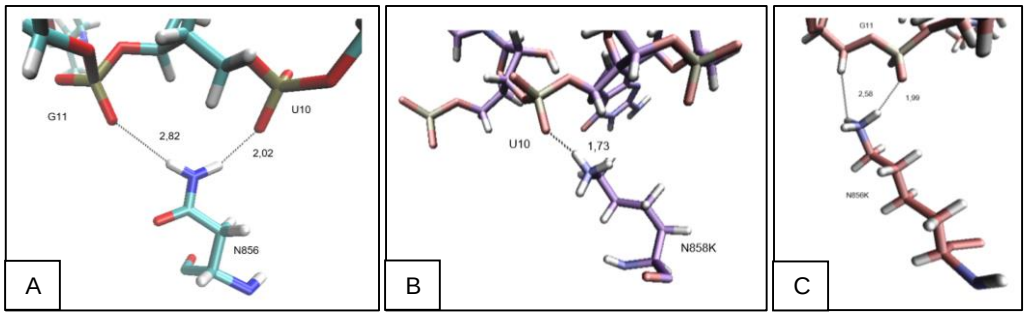


Figure 14 - Representation of the residue 856 of the (A) wild type, (B) N124K_N856K_V862K and (C) N856K structures and the RNA nucleotides U10 and G11. The values presented are in Ångström (Å).

In Figure 15 it is possible to see that is the backbone of the residue V862 which interacts with the nucleic acids, and that a change in its side chain (in the N124K_N856K_V862K (Figure 15B) and on V862K (Figure 15C)), does not infer with the interactions previously established nor with their number.

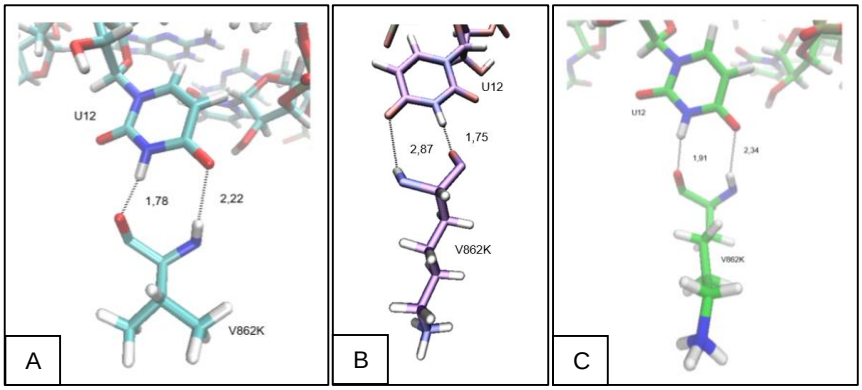


Figure 15 - Representation of the residue 862 of the (A) wild type, (B) N124K_N856K_V862K and (C) V862K structures and the RNA nucleotide U12. The values presented are in Ångström (Å).

Similar to what was done for the PAM section analysis, in this stage, it was also used RMSD and RMSF analysis to understand whether this mutation induced instability on the

enzyme, and the analysis were performed for the Cα residues of the protein, as is presented in Figure 16.

In this Figure, we can observe that the structure that was only mutated in the residue N856 presented the shortest estabilization period, and it also is the most stable structure from the four analysed. This is possible to see, not only through the RMSD analysis, but also from the RMSF analysis, where the N856K mutant presented one of the lowest number of spikes, with also the lowest values. In this structure, it easily noticed a spike for the residue ASP431, which is contained in an external loop of the REC2 domain, as well as a spike for ASP947, that also makes part of a loop, but this time of the Bridge Helix (BH) domain. Furthermore, the graphics for RMSF present some differences for N124K_N856K_V862K (C), where it is visible a spike for the residue GLN745, already described in the "PAM" section. As previously mentioned, this residue is present in one of the external loops of the protein and its unstability on the N124K_N856K_V862K mutant is surprising because this residue is less external in this mutant than in the wild type enzyme, being this last the structure where it was expected to present more variablility. Additionally, for the V862K mutant is visible a spike for the residue ILE150, located in an external loop of the REC1 domain.

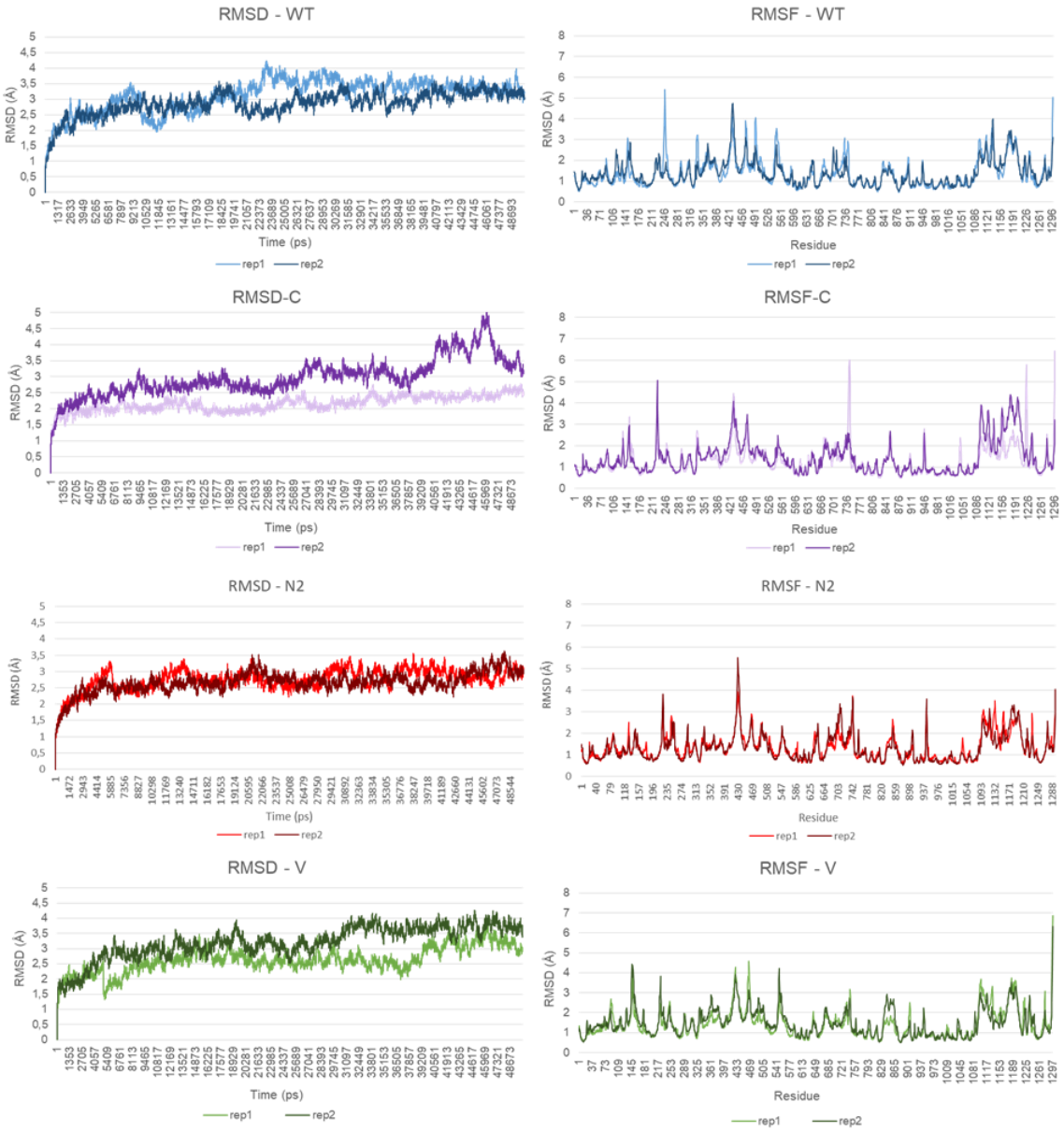


Figure 16 - Representation of Root Mean Square Deviations (RMSD) and Root Mean Square Fluctuations (RMSF) for the first and second replica of the wild type structure (WT) (rep1 and rep2, in light blue and dark blue, respectively), the N124K_N856K_V862K (C) (rep1 and rep2, in lilac and purple, respectively, as well), the mutant N856K mutant (N2) (rep1 and rep2, in red and brown, respectively) and of V862K (V) (rep1 and rep2, in light green and dark green, respectively).

In accordance to the analysis performed targeting the PAM section, in this one it was also decided to analyse the distances and angles between chosen atoms of the variants in analysis. This was done despite the knowledge that the exchange of our selected residues by a lysine would create an interaction charge-charge (stronger than an interaction by a hydrogen bond), as we were curious about how the distances between residues would

vary in comparison with the wild type structure. These results are presented in Figure 17, below.

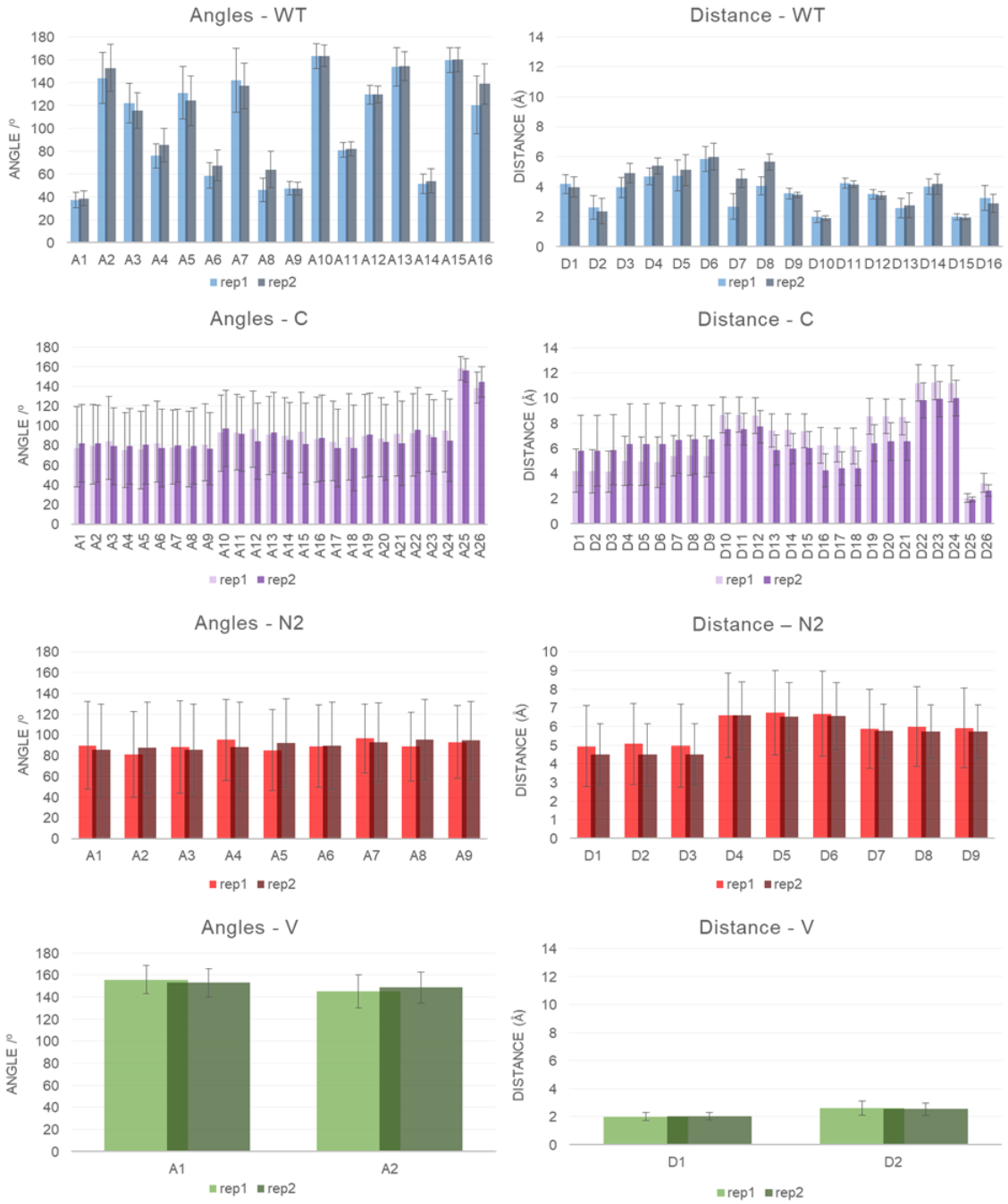


Figure 17 - Representation of mean values of the angles and distances, for the first and second replica of the wild type structure (rep1 and rep2, in light blue and dark blue, respectively), the N124K_N856K_V862K (C) (rep1 and rep2, in lilac and purple, respectively, as well) and the N856K mutant (N2) (rep1 and rep2, in red and brown, respectively) and of V862K (V) (rep1 and rep2, in light green and dark green, respectively).

As it is possible to distinguish when consulting the second section of Anex II - Part2, the values of N856K correspond to those from 9 to 14 of the WT structure for the distances and for the angles and to those from 10 to 24 of the N124K_N856K_V862K

mutant, for both the distances and the angles. For V862K, the values for both the distances and angles in the WT structure are the 15 and the 16, which correspond to the 25 and 26 on the N124K_N856K_V862K mutant.

In Figure 17, it is possible to observe that for the residue 856, in the WT structure, the best values for both the angles and the distance are presented by A10 and D10, as well as by A13 and D13, these values indicate the possible occurrence of hydrogen brigdes between the atom HD22 of the N856 and the OP1 of U10 (A10 and D10) and between HD21 of the N856 and the OP1 of G11 (A13 and D13). It is also possible to see that despite the interactions established, the mutant N124K_N856K_V862K presented higher distance values than the wild type structure for the residues analysed.

As we were able to observe from Figure 15, the variants for the residue 862 altered the side chain of this amino acid, however it was its backbone which interacted with the protein. Nonetheless, the distances and angles that this residue forms with the nucleotides of the crRNA 5'-handle were analysed. From this analysis, it was noticed that despite not being possible to see significative differences between the values for the angles and distances of the wild type structure and the structure of the N124K_N856K_V862K, the residue K862 is closer to the DNA and presents a lower deviation in the distances and angles measures, what may indicate that the mutation not only strengthened the interaction as well as made it more stable.

Enzyme mechanism

In order to understand which residues are important for transition state stabilization and are coordinating the magnesium ions, the literature was revised, enzymatic variants were developed and the structures were analysed. Additionally, two mutants were created for this study - presented in the section named "Catalytic site".

In this work, we put forward the hypothesis that Cas12a activity is coordinated by the residues D917, R918, E1006, H922, R1218 and D1255, being H922 the catalytic base (receiving the proton from the nucleophilic water molecule) (Figure 18). This hypothesis was formulated based on the following facts: it is known that in the catalysis of Cas9 - an enzyme of the same class of Cas12a - the residue that acts as the base is also a histidine (Casalino, *et al.*, 2019); it is possible to observe in Figure 14, that the mentioned residues are very close to the magnesium ions coordinating the cleavage, and so we suggest that D917, E1006 and D1255 make part of a DDE motif (a characteristic of two-metal aided

mechanism enzymes' architecture) - what also happens in Cas9, as demonstrated by Casalino, *et al*., who showed that RuvC catalytic site of this molecule and its magnesium ions are coordinated by E762, D986, D10 and S15 (Casalino, *et al*., 2019); In aqueous solution, a histidine has a pka of approximately 6. In our study, we did not calculate the pka for the histidine 922, but as we are studying the protein in a system with explicit solvent, we assumed that the pka for H922 would be close to this value, making it prone to protonation, since it has a free nitrogen atom and is aligned with the cleavage site and with what we propose that may be the nucleophilic water, captured on the crystallography.

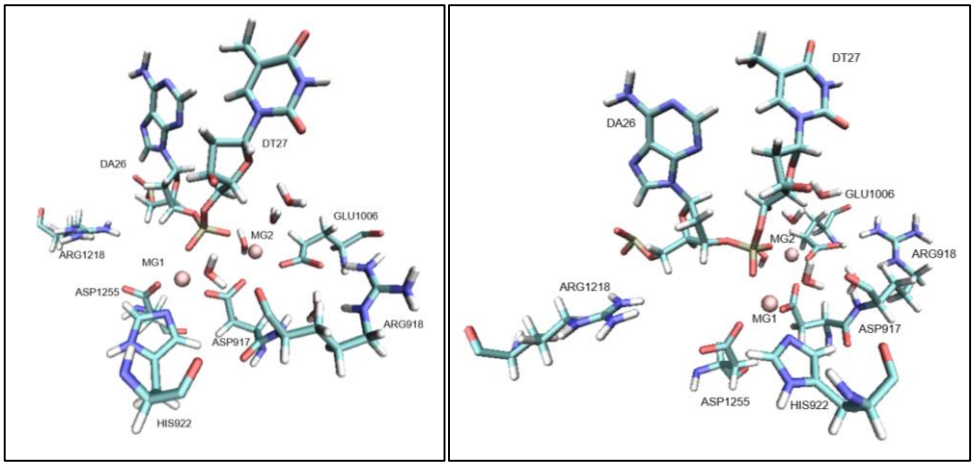


Figure 18 - Representation of different perspectives of the catalytic site of Cas12a.

It was already understood that the cleavage of the DNA phosphodiester bond contemplates several stages, which also encompass several conformational changes: the enzyme complexes with the crRNA, followed by the search for a complementary target DNA preceded by the PAM recognized by our enzyme. After this, there is the formation of a bubble when the DNA strands unzip, what leads to the cleavage of the non-template strand (Stella, *et al*., 2018). Here, we propose that for the DNA cleavage to occur, after the complex formation - with both DNA and RNA strands -, H922 is able to receive a proton from the nucleophilic water, as the water can attack the phosphorous atom of the phosphate group of DT27, which will receive the hydroxyl group formed on one cleaved nucleotide end and the hydrogen connected to the histidine on the other.

The study of this mechanism was performed by creating intermediate structures of the mechanism. A structure named PC1 was created, where H922 is protonated (Figure 19A) and a structure named PC2, where the target DNA is cleaved and H922 returned to its deprotonated state (Figure 19B and 19C). For the cleaved DNA structures, we had to

model two new units: a DNA 5' terminus with a phosphate group and the terminal 3' deprotonated, which are absent in common force fields. The atomic partial charges for the new units were calculated employing the (RESP) method (Bayly, *et al*., 1993) from the HF/6-31G(d,p) single-point energy calculations. Both structures were subjected to MD simulations, which were analysed afterwards.

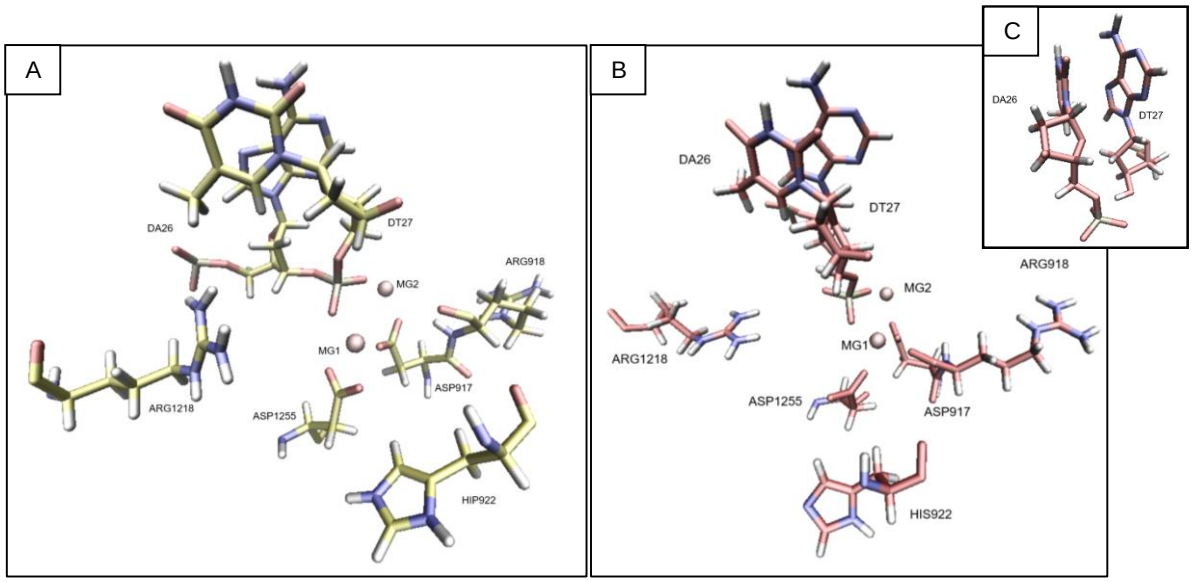


Figure 19 - Representation of the WT structure in the proposed states: A - WT enzyme with H922 protonated (HIP922); B - WT enzyme and nucleotides DA26 and DT27 cleaved (better observed in C, a rotation of DA26 and DT27 in B).

Through comparison between the structures obtained, as well as from a distance analysis (Figure 20), we can see that while H922 is aligned with the cleavage site during the WT structure (Figure 16) MD simulation, the same cannot be said for the PC1 and PC2 structures (Figure 19A and 19B, respectively), where H922 is further away from the catalytic site. These seems to indicate that if this histidine is effectively the catalytic residue, the proton transfer and nucleophilic attack might occur in a concerted mechanism, instead of step-wise one. This means that the proton transference between molecules can occur directly, instead of the structure transitioning through several intermediary states for said transference to take place or that other residue can act as the base

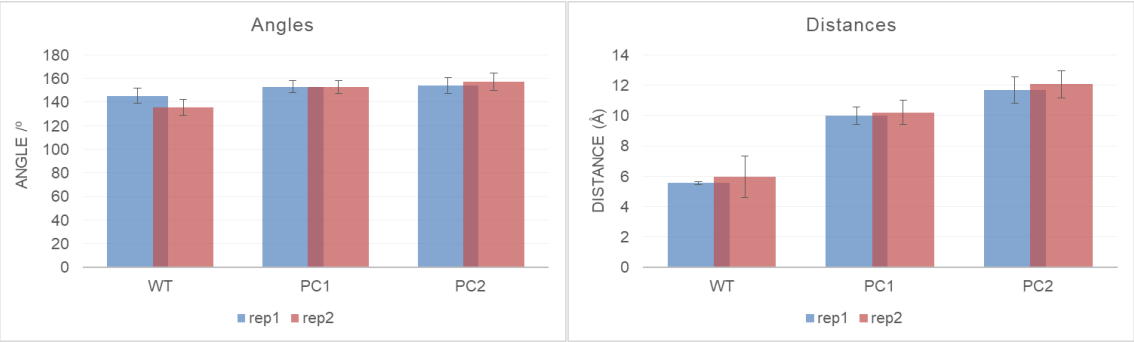


Figure 20 - Representation of mean values of the angles and distances, for the first (rep1 - blue) and second (rep2 - red) replica of the wild type structure (WT), the protonated structure (PC1) and the structure with the cleaved DNA (PC2).

From the Figure 20, it is clear that although in the WT structure the NE2 atom is aligned with the phosphate of DT27 and is relatively close to it, in the PC1 and PC2 structures, while the alignment between atoms is kept, the distances are higher. This result indicates that once H922 is protonated, it moves away from the catalytic site, what increases the difficulty to transfer the proton received from the nucleophilic water to the nucleic acid, cleaving the phosphodiester bond between DA26 and DT27.

To further study of this structures, it was also decided to analyse the MD simulations ran through RMSD and RMSF analysis (Figure 21). In the graphics made for the results obtained, we can observe that the intermediary structure with the cleaved DNA (PC2) only stabilized after approximately 27 nanoseconds of simulation. Furthermore, it is curious to notice that there is a spike present for the RMSF of the WT structure - for the residue TYR248 - that is missing on the PC1 and PC2 structures. This residue is part of a loop in the exterior of the protein - located in the REC1 domain - and is more external in the WT structure than in the others, what may explain this spike.

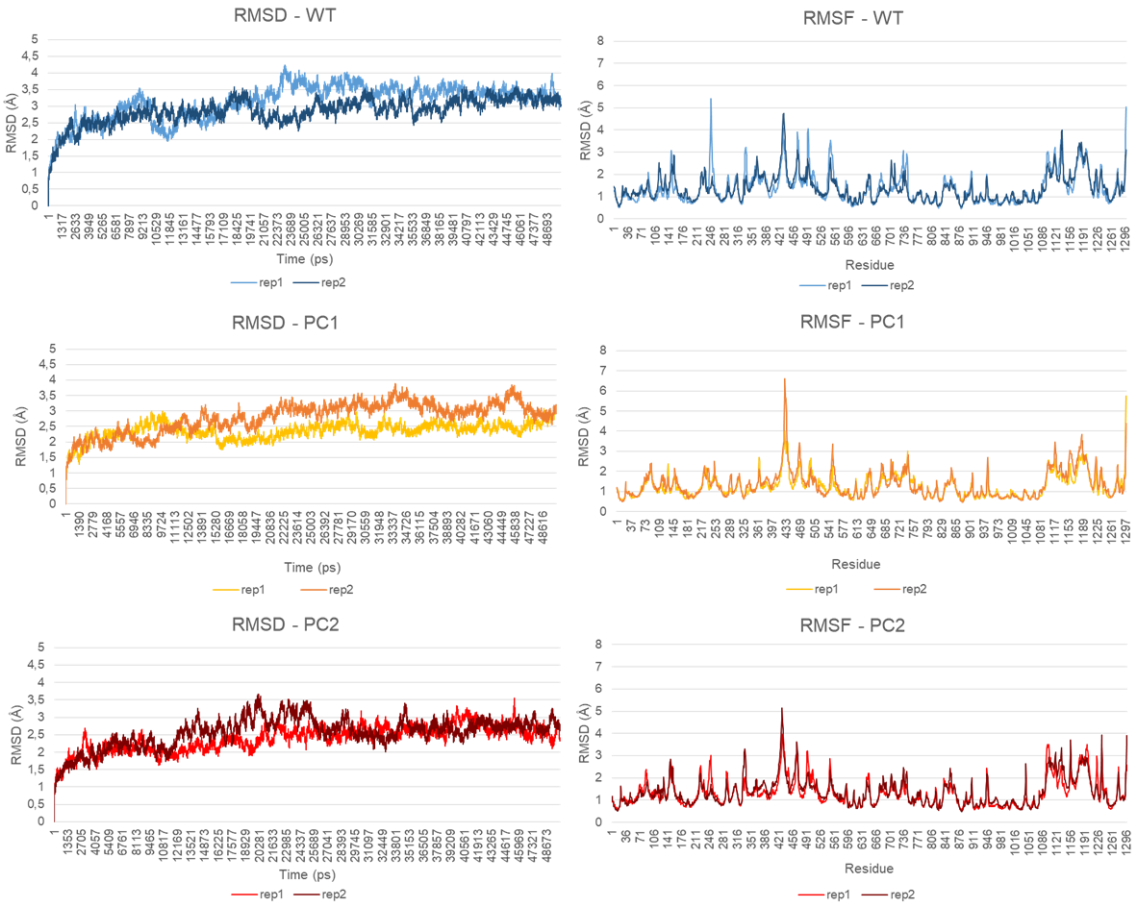


Figure 21 - Representation of Root Mean Square Deviations (RMSD) and Root Mean Square Fluctuations (RMSF) for the first and second replica of the wild type structure (WT) (rep1 and rep2, in light blue and dark blue, respectively), the protonated histidine structure (PC1) (rep1 and rep2, in yellow and orange, respectively as well) and the structure with the cleaved DNA (PC2) (rep1 and rep2, in red and brown, respectively).

To deepen the study of the catalytic mechanism of this enzyme, two mutants were created, H922A and R1218K (Figure 22A and 23A, respectively) which would potentially offer more insight on this enzyme activity. With H922A we tried to validate the hypothesis that H922 is the catalytic residue and that without it the reaction does not occur - it would be an inactive form of the enzyme. R1218K was created with the intention of better understanding the relevance of this residue and to study whether a lysine residue would increase the stability of the magnesium ions in the active site of the molecule, promoting the catalysis.

For this purpose, the selected residues were exchanged, their geometry optimized through the previously described protocol and 50 nanoseconds MD simulations performed.

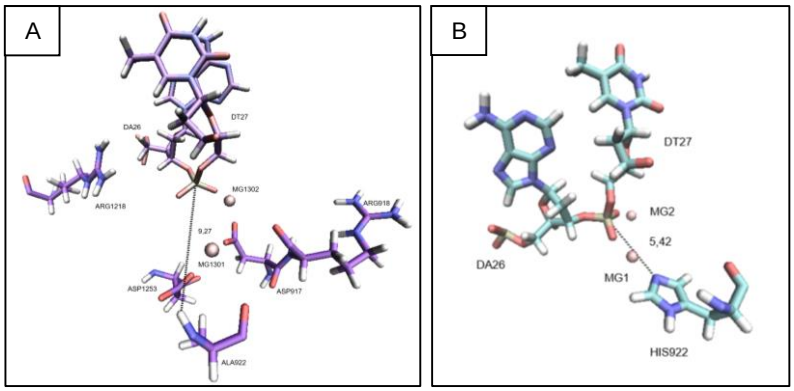


Figure 22 - Representation of the active site of the mutant H922A (A), with focus on the residue 922. B) - interactions established by H922 in the WT enzyme with the DNA nucleotides DA26 and DT27. The values presented are in Ångström (Å).

In Figure 22, it is possible to observe that when the H922 is exchanged by an alanine, the residue moves away from the catalytic site.

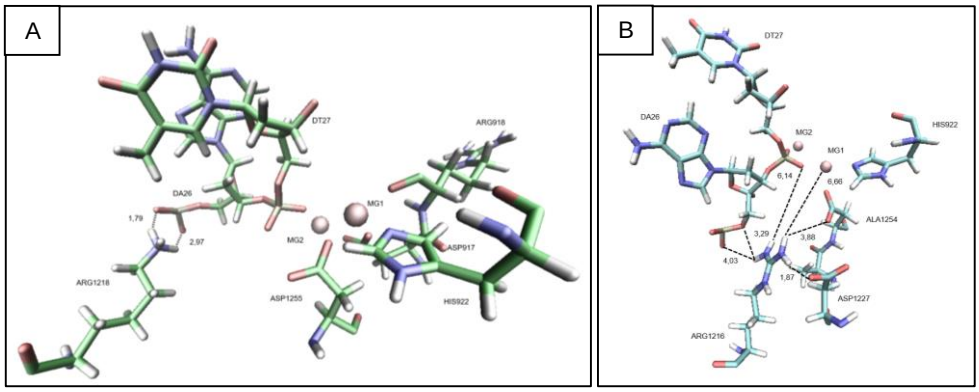


Figure 23 - Representation of the active site of the mutant R1218K (A) with focus on the residue R1218K. B) - interactions established by R1218 in the WT enzyme. The values presented are in Ångström (Å).

From the Figure 23, it is possible to observe that the R1218K mutant (Figure 23A) shifted the interactions established by the R1218 residue in the WT enzyme (Figure 23B).

To better understand the impact of these mutations, RMSD and RMSF analysis was performed, as presented in Figure 24 below.

From the RMSD analysis, it is possible to observe that the structures stabilize fairly quickly. Despite that, there are several spikes in the RMSF analysis that catch our attention, such as the spike at the residue TYR248 of the wild type structure (that as previously referred is present in an external loop of the REC1 domain - that is absent in the R1218K and H922A mutants; the value of the spike for the residue ASP431 - also present in one of the molecule' external loops, in the REC2 domain - is higher in the H922A mutant than in the other structures and the spike on the residue ARG946 on the

R1218K structure, what is curious, as this residue is part of the Bridge Helix domain, which is central in the protein and it is relatively close to the crRNA 5' handle.

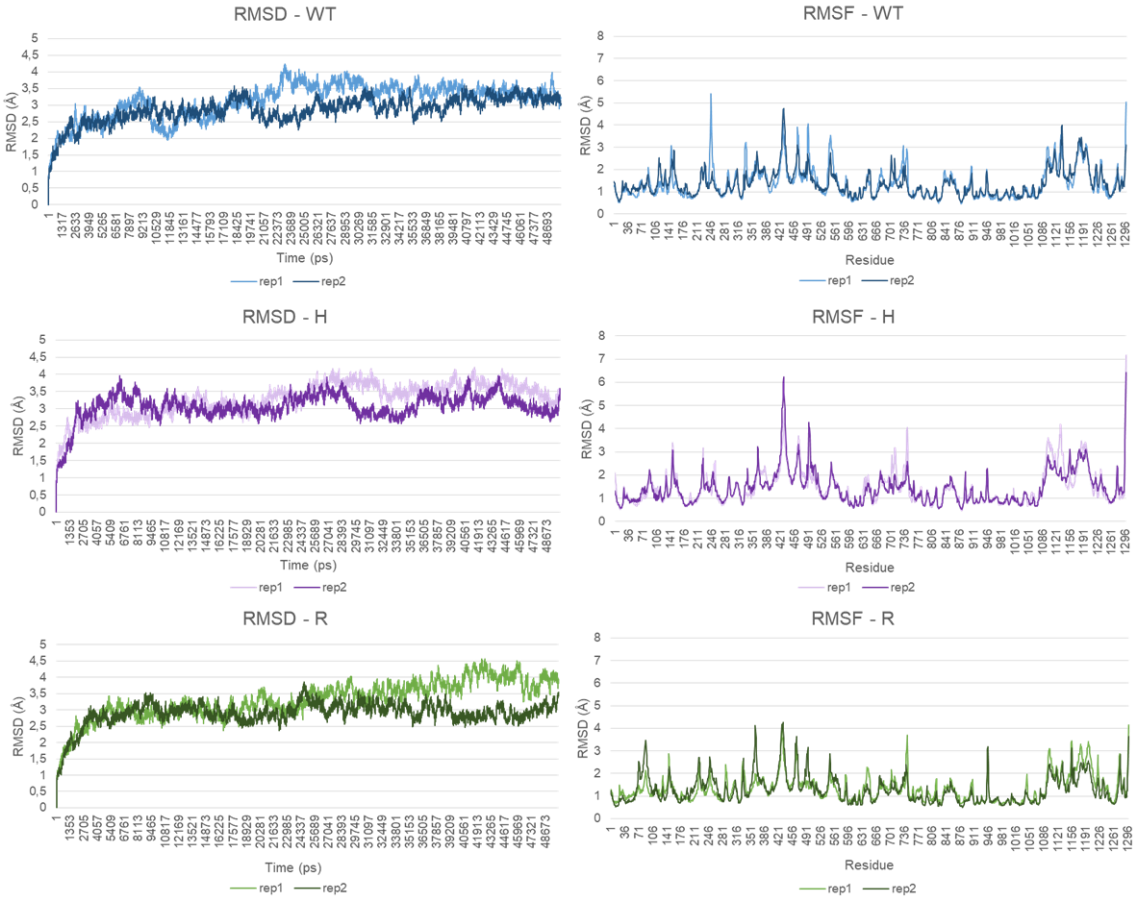


Figure 24 - Representation of Root Mean Square Deviations (RMSD) and Root Mean Square Fluctuations (RMSF) for the first and second replica of the wild type structure (WT) (rep1 and rep2, in light blue and dark blue, respectively), the mutant H922K (H) (rep1 and rep2, in lilac and purple, respectively as well) and the structure R1216K (R) (rep1 and rep2, in light green and dark green, respectively).

QM/MM

After understanding that the catalytic activity of this enzyme may occur through a concerted mechanism, it was decided to proceed with a QM/MM study of the catalytic region. For this, we selected to use the WT structure to study the mechanism in a forward fashion and the PC2 structure to attempt to analyze the mechanism in reverse.

Different atom combinations were selected for the high layer over several attempts, however we were not able to advance with the calculations and we formulated the hypothesis that the atoms we were choosing were either wrong and/or too distant for the performance of a QM/MM study. To surpass this obstacle, we decided to bring H922

closer to the cleavage site, what was done by applying forces to the NE2 atom of H922 and to the phosphorus atom of DT27, with a strength of 500 kcal/mol/Å, to bringing this residues to a distance of 3 Å on the wild type structure (from a distance of approximately 6,2 Å) and to a distance of 5 Å on the intermediate structure with the cleaved DNA (from a distance of approximately 10,8 Å).

These approximations were concluded successfully, however, after a 50 nanosecond MD simulation, it was observed that the PC2 structure returned to a distance of approximately 11 Å, and so, it was decided to abandon this strategy and solely proceed with the wild type structure.

Due to time constraints, it was not possible to properly understand and describe the mechanism of this endonuclease. Nevertheless, we are left with several starting points to further advance the knowledge that has already been acquired.

MMPBSA

To perform a MMPBSA analysis it was necessary to create the necessary structures to proceed with the analysis. These consisted on the unsolvated structures for the ligand - in our case, the nucleic acids, the receptor - which consisted in our variants, and the complex - the ribonucleoprotein. These structures allow the calculation of the binding free energy of the structure through the thermodynamic cycle.

For the preparation of these structures, it was decided to erase every water molecule from the pdb files (to correctly simulate the in vacuum system) as well as the final nucleotides of each nucleic acid chain (to reduce potential instability), before creating the topology files needed for the analysis - which was performed using the default values for the MMPBSA analysis with perl.

Due to time constraints, it was only possible to present the values obtained for one of the replicas of each enzymatic variant created, these are presented in Figure 25, below.

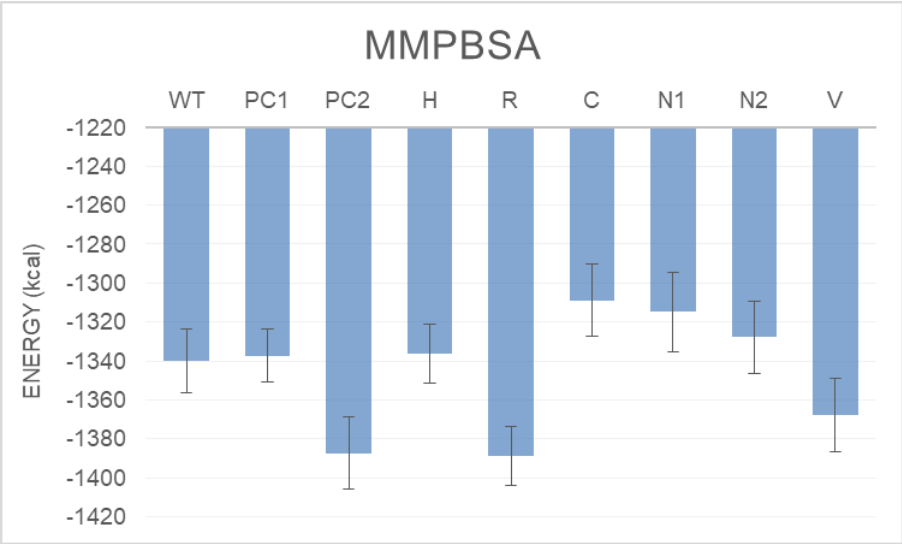


Figure 25 - Values obtained for the MMPBSA analysis for the 9 enzimatic variants studied during this project: wild type structure (WT), first intermediate structure (PC1), second intermediate structure (PC2), H922A mutant (H), R1216K mutant (R), N124K_N856K_V862K (C), N1224K (N1), N856K (N2), V862K (V).

From this analysis we can observe that the values obtained are very different among them, what was not expected, but may be explained by the elevated flexibility of some protein motifs and may as well indicate that the structures need a longer simulation time in order to further stabilize. This result may also be a reflexion of the modelling of a long stretch of the DNA non-template strand that was absent in the initial structure.

Wet Lab Results and Discussion

Concentration of plasmid DNA stocks:

This evaluation was performed using a Nanodrop ND-1000 Spectrophotometer (Thermo Scientific). This measures the concentration of plasmid DNA of the obtained samples, allowing the progression of the work to the next stage: cell transformation with our plasmid of interest.

The results obtained were:

Replica 1 - 96,5 ng/µL; Replica 2 - 125,5 ng/µL (average of 111 ng/µL)

This was according to the expected (~100 ng/µL) when using the NZYTech purification kit and the conditions described before.

Bacterial transformation:

By transforming *E. coli* BL21 Star (DE3) with the plasmid pMBP-FnCas12a we intended to obtain enough cellular material to create stock solutions containing our plasmid of interest. This step enables us to save time posteriorly, as well as to proceed to the expression of Cas12a, a first step to validate this protocol and advance to the testing the mutant enzyme developed computationally.

As it is possible to observe in the Figure 26, this step was performed successfully.

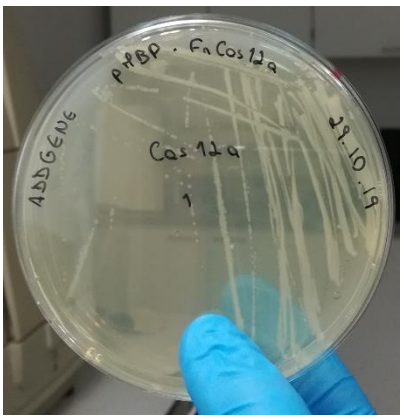


Figure 26 - Plate containing colonies of *E. coli* BL21 Star (DE3) transformed with the plasmid pMBP-FnCas12a - Image captured by Ana Teresa Rajado on the 29th October 2019.

Sequencing:

The results of the sequencing at GATC Services - Eurofins Genomics (Germany) were analysed through comparison between the sequences obtained from our samples and the sequence of the plasmid pMBP-FnCas12a received from Addgene. It was possible to observe that both primers were able to sequence our fusion protein - the T7 Promoter primer allowed the transcription of the beginning of MBP and the T7 Terminator primer allowed the transcription of the end of FnCas12a. The fact that the 3' and 5' ends of our fusion protein were present, gave us confidence to assume that the plasmid received contained our whole sequence of interest.

Protein Expression Evaluation:

We used two parallel approaches to evaluate the protein expression. Pre-cultures were prepared by using two samples of transformed from the glycerol cell stocks previously kept at - 80 ºC and two other samples were freshly transformed from the plasmid DNA stock kept at -20ºC. These samples were transformed and with pMBP-FnCas12a and the protein expression was induced in accordance to the protocol.

To evaluate whether the expression of our protein of interest was successful, an SDS-PAGE electrophoresis on polyacrylamide gel was performed. Two samples of each replica were collected, one before the expression induction with IPTG (T0) and the other sample 16 hour after (T_{ON}) and their OD measured (Table 2).

Table 2 - Optical Density (OD) values of the solutions used to evaluate the protein' expression. T0 indicates the values of the bacterial solution with transformed cells before inducting protein expression and T_{ON} presents the OD values 16 hours after induction with IPTG.

Replica	T0	T_{ON}
1	0,69	2,92
2	1,01	3,28
3	0,92	3,32
4	0,78	3,36

By collecting and using these samples, we expected to observe significant differences among them: while in the T0 samples we expected to observe thin bands of the proteins contained in the cells, in the T_{ON} sample we expected being able to distinguish a bold

band at approximately 197 kDA, corresponding to MBP-FnCas12a. As it is possible to observe in the Figure 27, this didn't happen, and we did not observe a clear overexpressed band of MBP-FnCas12a. This might have happened for several motives, such as: we could have received a different plasmid sample from Addgene than what we asked for, but we were able to demonstrate that this as not the case through sequencing; although we were very attentive, it is not possible to discard the possibility of cross contamination from either the reagents or appliances used; moreover, as we used reagent that were already in the lab, it is possible that some old stock solutions were not optimal. In addition, even though the protocol used was reliable and already used in other studies, it might need further optimizations in order to be applied to the conditions of the host lab.

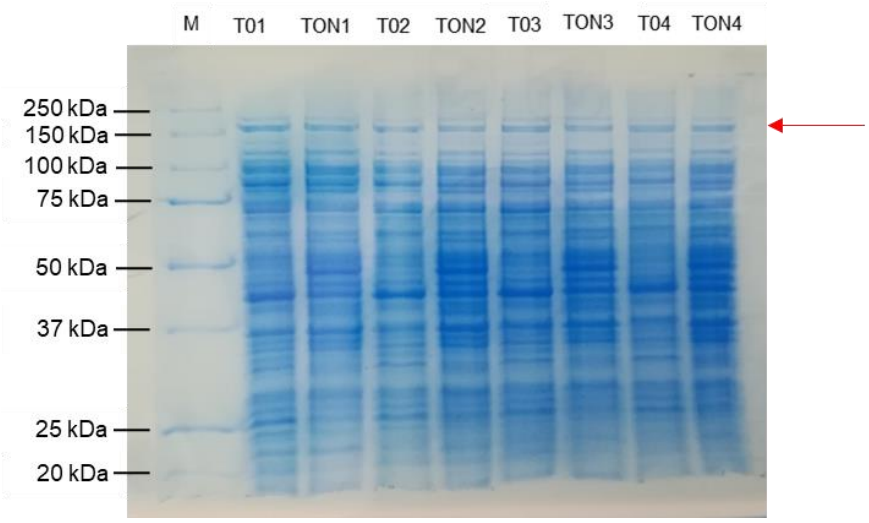


Figure 27 - 12% SDS-PAGE gel, stained with Coomassie Blue, used to evaluate MBP-FnCas12a expression - Image captured by Ana Teresa Rajado on the 6th march 2020. Legend: M - Marker; T0 - Protein expression before induction; TON - Protein expression 16h after induction. The red arrow indicates 197 kDa, the place where our fusion protein was expected to be in.

Uncompleted tasks

This section presents the two stages planned at the beginning of this work which were not executed due to Sars-CoV-2 pandemic. These stages are based on the "Heterologous Expression and Purification of CRISPR-Cas12a/Cpf1" protocol (Mohanraju, et al., 2018), being the first stage the only one explored on this work.

- Stage 2 - the enzyme would be purified in a two-step process: 1) a first purification step making use of an affinity chromatography with HisTrap columns (GE Healthcare)

and an ÄKTA FPLC system, followed by another 2) affinity chromatography with an Heparin FF column. The first will allow purification through the N-terminal 10His tag and the second one will make use of the Heparin ability to bind DNA-binding proteins. Between both chromatographies a TEV cleavage step will enable removal of the MBP tag.

- Stage 3 - this stage would enable a comparative evaluation of the cutting efficiency between the recombinant FnCas12a wild type version and a mutant version to be prepared by site-directed mutagenesis, mutating critical residues involved in the catalytic function of the enzyme. An in vitro cleavage assay making use of a well characterized plasmid DNA as target will help elucidating the preliminary cutting efficiency between WT and mutant. We anticipate that a rather more refined assay will need to be considered/developed in order to adequately quantify cutting efficiency rate between both (Mohanraju, et al., 2018).

Conclusions and Future Perspectives

This project opens the door to the study of the catalytic mechanism of FnCas12a. In here, we put forwards several hypothesis to explore different pathways still unsolved and faced trials that require the formulation of new solutions. Although many of questions remain unanswered, we were able to bring some light to this topic:

- We were able to understand that the interactions we exchanged (from hydrogen bonds to salt bridges) lead to stronger connections between the protein and both nucleic acids it complexes with. While it is still uncertain whether any of our variations increased the specificity of the enzyme in any of our mutants, this question may be answered once further analysis, such as the MMPBSA, is performed and a model validated in the wet lab;
- With this work, we were able to propose H922 as an important residue for the catalytic mechanism;
- The comparisons between the intermediate structures for the catalytic mechanism of this enzyme suggest us that the mechanism might occur by a concerted mechanism;
- We were able to understand that despite our model requiring further optimization, the work accomplished left us several starting points from which to proceed.

In a world which constantly requires more and better from its citizens to face the adversities - as this pandemic came to prove - it has become increasingly more relevant to understand what surrounds us. While that still includes trying to understand the mechanism behing this enzyme catalytic activity and how to reduce its off-target mutations, it is our hope that this work may clarify some points and perhaps be the inspiration for new projects.

References

Aehle, W. (2007) *Enzymes in Industry*. Wiley-VHC

Almeida, B. C., Figueiredo, P., & Carvalho, A. T. P. (2019). Polycaprolactone Enzymatic Hydrolysis: A Mechanistic Study. *ACS Omega*, *4*(4), 6769–6774. https://doi.org/10.1021/acsomega.9b00345

Aouida, M., Eid, A., Ali, Z., Cradick, T., Lee, C., Deshmukh, H., Atef, A., AbuSamra, D., Gadhoum, S. Z., Merzaban, J., & Mahfouz, M. (2015). Efficient fdCas9 synthetic endonuclease with improved specificity for precise genome engineering. *PLoS ONE*, *10*(7), 1–16. https://doi.org/10.1371/journal.pone.0133373

Arab, S. S., Shirvanizadeh, N., & Khoshnejat, M. (2014). A collection of bioinformatics tools to protein engineering. *Molecular Biology Research Communications 3*, 32–77.

Arodola, O. A., & Soliman, M. E. S. (2016). Molecular Dynamics Simulations of Ligand-Induced Flap Conformational Changes in Cathepsin-D-A Comparative Study. *Journal of Cellular Biochemistry, 999*, 1-15.

Bae S., Park J., & Kim J. S. (2014) Cas-OFFinder: A fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics, 30,* 1473-1475. https://doi.org/10.1093/bioinformatics/btu048

Bakoweis, D., & Thiel, W. (1996). Hybrid Models for Combined Quantum Mechanical and Molecular Mechanical Approaches. *J. Phys. Chem. 100,* 10580-10594

Bartlett, D. F., Goldhagen, P. E. & Phillips, E. A. (1970). Experimental Test of Coulomb's Law. *Physical Review*, *2*(2), 483-487. https://doi.org/10.1103/PhysRevD.2.483

Bayly, C. I., Cieplak, P., Cornwell, W. D., Kollman, P. A. (1993). A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model. *Journal of Physical Chemistry, 97*, 10269-10280. https://doi.org/10.1021/j100142a004

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research, 28*, 235-242. https://doi.org/10.1093/nar/28.1.235

Blake, C. C., Koenig, D. F., Mair, G. A., North, A. C. T., Phillips, D. C., & Sarma, V. R. (1965). Structure of hen egg-white lysozyme. *Nature*, *206*(4986), 757–761.

Bogdanove, A. J., Bohm, A., Miller, J. C., Morgan, R. D., & Stoddard, B. L. (2018). Engineering altered protein–DNA recognition specificity. *Nucleic Acids Research*, *46*(10), 4845–4871. https://doi.org/10.1093/nar/gky289

Bornscheuer, U. T., & Pohl, M. (2001). Improved biocatalysts by directed evolution and rational protein design. *Current Opinion in Chemical Biology*, *5*(2), 137–143. https://doi.org/10.1016/S1367-5931(00)00182-4

Brannigan, J. A., & Wilkinson, A. J. (2002). Protein engineering 20 years on. *Nature Reviews Molecular Cell Biology*, *3*(12), 964–970. https://doi.org/10.1038/nrm975

Carvalho, A. T. P., Barrozo, A., Doron, D., Vardi, K. A., Major, D. T., & Kamerlin, S. C. L. (2014). Challenges in computational studies of enzyme structure, function and dynamics. *Journal of Molecular Graphics and Modellin*g, *54*, 62–79. https://doi.org/10.1016/j.jmgm.2014.09.003

Casalino, L., Jinek, M., & Palermo, G. (2019). Two-metal ion mechanism of DNA cleavage by CRISPR-Cas9. ChemRxiv. Preprint. https://doi.org/10.26434/chemrxiv.9784082.v1

Case, D. A., Ben-Shalom, I. Y., Brozell, S. R., Cerutti, D. S., Cheatham, T. E., III, Cruzeiro, V. W. D., Darden, T. A., Duke, R. E., Ghoreishi, D., Gilson, M. K., Gohlke, H., Goetz, A. W., Greene, D., Harris, R., Homeyer, N., Izadi, S., Kovalenko, A., Kurtzman, T., Lee, T. S., LeGrand, S., Li ,P., Lin, C., Liu, J., Luchko, T., Luo, R., Mermelstein, D. J., Merz, K. M., Miao, Y., Monard, G., Nguyen, C., Nguyen, H., Omelyan, I., Onufriev, A., Pan, F., Qi, R., Roe, D. R., Roitberg, A., Sagui, C., Schott-Verdugo, S., Shen, J., Simmerling, C. L., Smith, J., Salomon-Ferrer, R., Swails, J., Walker, R. C., Wang, J., Wei, H., Wolf, R. M., Wu, X., Xiao, L., York, D. M., & Kollman P. A. (2018). AMBER 2018, University of California, San Francisco.

Caulfield, B. (2009). *NVIDIA - What's the Difference Between a CPU and a GPU?* Retrieved on the 30th of December of 2019 from https://blogs.nvidia.com/blog/2009/12/16/whats-the-difference-between-a-cpu-and-a-gpu/

Chandrasegaran, S., & Carroll, D. (2016). Origins of Programmable Nucleases for Genome Engineering. *Journal of Molecular Biology*, *428*(5), 963–989. https://doi.org/10.1016/j.jmb.2015.10.014

Christensen, A. S., Kubar, T., Cui, Q., & Elstner, M. (2016) Semiempirical Quantum Mechanical Methods for Noncovalent Interactions for Chemical and Biochemical Applications. *Chemical Reviews, 116*, 5301-5337.

Cyranoski, D. & Ledford, H. (2018) Genome-edited baby claim provokes international outcry. *Nature,* 563, 607-608, doi: 10.1038/d41586-018-07545-0

Dong, D., Ren, K., Qiu, X., Zheng, J., Guo, M., Guan, X., Liu, H., Li, N., Zhang, B., Yang, D., Ma, C., Wang, S., Wu, D., Ma, Y., Fan, S., Wang, J., Gao, N., & Huang, Z. (2016). The crystal structure of Cpf1 in complex with CRISPR RNA. *Nature, 532,* 522–526. https://doi.org/10.1038/nature17944

Dourado, Daniel (2010*) Glutathione transferase: theoretical insights*. Departamento de Química - Faculdade de Ciências da Universidade do Porto.

Durrant, J. D., & McCammon, J. A. (2011). Molecular dynamics simulations and drug discovery. *BMC Biology, 9,* 71. doi:10.1186/1741-7007-9-71

Facciotti, M. T., Bertain, P. B. & Yuan, L. (1999). Improved stearate phenotype in transgenic canola expressing a modified acyl–acyl carrier protein thioesterase. *Nature Biotechnology, 17*(6), 593–597

Fonfara, I., Richter, H., Bratovič, M., Rhun, A. Le, & Charpentier, E. (2016). The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. *Nature,* 532, 517–521. https://doi.org/10.1038/nature17945

Friesner, R. A., & Guallar, V. (2005). Ab Initio Quantum Chemical and Mixed Quantum Mechanics/Molecular Mechanics (Qm/Mm) Methods for Studying Enzymatic Catalysis. *Annual Review of Physical Chemistry, 56*(1), 389–427. https://doi.org/10.1146/annurev.physchem.55.091602. 094410

Fuxreiter, M. (2014). *Computational Approaches to Protein Dynamics: From Quantum to Coarse-Grained Methods*. Boca Raton: Taylor &Francis Group

Galindo-Murillo, R., Robertson, J. C., Zgarbová, M., Šponer, J., Otyepka, M., Jurečka, P., & Cheatham, T. E. (2016). Assessing the Current State of AMBER Force Field Modifications for DNA. *Journal of Chemical Theory and Computation, 12*, 4114-4127. DOI: 10.1021/acs.jctc.6b00186

Gao, P., Yang, H., Rajashankar, K. R., Huang, Z., & Patel, D. J. (2016). Type V CRISPR-Cas Cpf1 endonuclease employs a unique mechanism for crRNA-mediated target DNA recognition. *Cell Research, 26*(8), 901–913. https://doi.org/10.1038/cr.2016.88

Gao, L., Cox, D. B. T., Yan, W. X., Manteiga, J. C., Schneider, M. W., Yamano, T., Nishimasu, H., Nureki, O., Crosetto, N., & Zhang, F. (2017). Engineered Cpf1 variants with altered PAM specificities increase genome targeting range. *Nature Biotechnology*, *35*(8), 789–792. https://doi.org/10.1038/nbt.3900

Gromacs Tutorial, (n.d.) *Analysis*. Retrieved on the 30th October 2020 from http://www.strodel.info/index_files/lecture/html/analysis-4.html

Harris, J. L., & Craik, C. S. (1998). Engineering enzyme specificity. *San Francisco: Current Opinion in Chemical Biology*, *2*(1), 127–132. https://doi.org/10.1016/S1367-5931(98)80044-6

Haas, Kathryn. (2019). *Coulomb's Law*. Retrieved on the 5th february of 2020 from https://chem.libretexts.org/Courses/Saint_Mary's_College%2C_Notre_Dame%2C_IN/CHEM_342%3A_Bioinorganic_Chemistry/Readings/Week_1%3A_Analysis_of_Periodic_Trends/1.1%3A_Concepts_and_principles_that_explain_periodic_trends/1.1.1%3A_Coulomb's_Law

Heler, R., Marrafini, L. A., & Bikard, D. (2013). Adapting to new threats: the generation of memory by CRISPR-Cas immune systems. *Molecular Microbiology*, *93*(1), 1–9. https://doi.org/10.1111/mmi.12640

Hille, F., & Charpentier, E. (2016). CRISPR-cas: Biology, mechanisms and relevance. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1707), 1239-1259. https://doi.org/10.1098/rstb.2015.0496

Hsu, P. (2014). Development and Applications of CRISPR-Cas9 for Genome Engineering. *Cell* 157(6), 1262–1278. https://doi.org/10.1016/j.cell.2014.05.010.

Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD - Visual Molecular Dynamics. *Journal of Molecular Graphics*, *14*, 33-38

Intel. (n.d.) *CPU vs. GPU: Making the most of both.* Retrieved on the 30th December of 2019 from https://www.intel.com/content/www/us/en/products/docs/processors/cpu-vs-gpu.html

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science*, *337*(6096), 816–821. https://doi.org/10.1126/science.1138140

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., & Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 926(May 2012). https://doi.org/10.1063/1.445869

Kleinstiver, B. P., Tsai, S. Q., Prew, M. S., Nguyen, N. T., Welch, M. M., Lopez, J. M., McCaw, Z. R., Aryee, M. J., & Joung, J. K. (2016). Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells. *Nature Biotechnology*, *34*(8), 869–874. https://doi.org/10.1038/nbt.3620

Kosloff, R. (1988). Time-Dependent Quantum-Mechanical Methods for Molecular Dynamics. *The Journal of Physical Chemistry* 92(19), 2087–2100.

Kuzmanic, A., & Zagrovic, B. (2010). Determination of Ensemble-Average Pairwise Root Mean-Square Deviation from Experimental B-Factors. *Biophysical Journal*, *98*(5), 861–871. https://doi.org/10.1016/j.bpj.2009.11.011

Mohanraju, P., van der Oost, J., Jinek, M., & Swarts, D. C. (2018). Heterologous expression and purification of CRISPR-Cas12a/Cpf1. *Bio-Protocol*. Stippeneng, Wageningen: Wageningen University. Zurich, Switzerland: University of Zurich. Retrieved from www.bio-protocol.org/e2842

Leach, A. R. (2001). *Molecular Modeling: Principles and Applications* (2nd ed.). In P. Hall (Ed.), Glaxo Wellcome Research and Development. https://doi.org/10.1021/ci9804241

Levine, I. N. (2013). *Quantum chemistry* (7th ed.). Upper Saddle River, NJ: Pearson Education.

Levitt, M., & Warshel, A. (1975) Computer simulation of protein folding. *Nature 253* 694–698

Levitt, M., Hirshberg, M., Sharon, R., & Dagget, V. (1995). Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. *Computer Physics Communications*, *91*, 215-231. https://doi.org/10.1016/0010-4655(95)00049-L

Li, B., Zhao, W., Luo, X., Zhang, X., Li, C., Zeng, C., & Dong, Y. (2017). Engineering CRISPR-Cpf1 crRNAs and mRNAs to maximize genome editing efficiency. *Journal of Autism and Developmental Disorders*, *47*(3), 549–562. https://doi.org/10.1097/CCM.0b013e31823da96d.Hydrogen

Lippow, S. M., & Tidor, B. (2007). Progress in computational protein design. *Current Opinion in Biotechnology*, *18*(4), 305–311. https://doi.org/10.1016/j.copbio.2007.04.00

Madej, B. D., Walker, R., Munshi, A., Mannan S., Barton, M., Luchko, T. (2015; 2018). *An Introduction to Molecular Dynamics Simulations using AMBER*. Retrieved on the 30th january of 2020 form http://AMBERmd.org/tutorials/basic/tutorial1/section5.htm

Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K., & Simmerling, C. (2015). ff14SB : Improving the accuracy of protein side chain and backbone parameters from ff99SB. *Journal of Chemical Theory and Computation*, *11*(8), 3696–3713.

Melo, M. C. R., Bernardi, R. C., Rudack, T., Scheurer, M., Riplinger, C., Phillips, J. C., Maia, J. D. C., Rocha, G. B., Ribeiro, J. V., Stone, J. E., Neese, F., Schulten K., & Luthey-Schulten, Z. (2018). NAMD goes quantum: An integrative suite for QM/MM simulations. *Nature Methods* *15*(5), 351–354. https://doi.org/10.1038/nmeth.4638.NAMD

Naeem, Rabia. (2019). *Lennard-Jones Potential*. Retrieved on the 5th february of 2020 from https://chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_Maps/Supplemental_Modules_(Physical_and_Theoretical_Chemistry)/Physical_Properties_of_Matter/Atomic_and_Molecular_Properties/Intermolecular_Forces/Specific_Interactions/Lennard-Jones_Potential

Nelson, David L., & Cox, Michael, (David Lee), 1942-. (2005). *Lehninger principles of biochemistry*. New York: W.H. Freeman

Nishimasu, H., Cong, L., Yan, W. X., Ran, F. A., Zetsche, B., Li, Y., Kurabayashi, A., Ishitani, R., Zhang, F., & Nureki, O. (2015). Crystal Structure of Staphylococcus aureus Cas9. *Cell*, *162*(5), 1113–1126. https://doi.org/10.1016/j.cell.2015.08.007

Nishimasu, H., Yamano, T., Gao, L., Zhang, F., Ishitani, R., & Nureki, O. (2017). Structural Basis for the Altered PAM Recognition by Engineered CRISPR-Cpf1. *Molecular Cell*, *67*, 139-147. https://doi.org/10.1016/j.molcel.2017.04.019

Palermo, G. (2019). Structure and Dynamics of the CRISPR-Cas9 Catalytic Complex. *Journal of Chemical Information and Modelling*. doi:10.1021/acs.jcim.8b00988

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., Ferrin, T. E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*. (13):1605-12. DOI: 10.1002/jcc.20084

Purves, K. W., Sadava, D., Orians, G. H., & Heller, H. C., (2004). *Life - the science of biology - Volume 1* (7th edition). Sunderland, Mass Sinauer Associates.

Ramos, R. M. C. (2012). *Determinação computacional dos efeitos da mutagénese em interfaces proteína-ADN*. Faculdade de Ciências da Universidade do Porto, https://repositorioaberto.up.pt/bitstream/10216/65143/2/24220.pdf

Razin, S. V., Borunova, V. V., Maksimenko, O. G., & Kantidze, O. L. (2012). Cys2His2 zinc finger protein family: Classification, functions, and major members. *Biochemistry (Moscow)*, *77*(3), 217–226. https://doi.org/10.1134/s0006297912030017

RCSB (n.d.). *Protein Data Bank*. Retrieved on the 27th January 2020, from https://www.rcsb.org/

RCSB - 6GTF (n.d.) *Protein Data Bank.* Retrieved on the 6th July 2019, from https://www.rcsb.org/structure/6GTF

Salomon-Ferrer, R., Case, D. A., & Walker, R. C. (2013). An overview of the AMBER biomolecular simulation package. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *3*(2), 198–210. https://doi.org/10.1002/wcms.1121

Salomon-Ferrer, R., Goetz, A. W., Poole, D., Grand, S. Le, & Walker, R. C. (2013). Routine microsecond molecular dynamics simulations with Routine microsecond molecular dynamics simulations with AMBER on GPUs . 2 . Explicit Solvent Particle Mesh Ewald. *Journal of Chemical Theory and Computation*, *9*(9), 3878–3888. https://doi.org/10.1021/ ct400314y

Souza, O. N., & Ornstein, R. L. (1997). Effect of Periodic Box Size on Aqueous Molecular Dynamics Simulation of a DNA Dodecamer with Particle-Mesh Ewald Method. *Biophysical Journal*, *2*, 2395-2397. doi:10.1016/S0006-3495(97)78884-2

Stella, S., Mesa, P., Thomsen, J., Paul, B., Alcón, P., Jensen, S. B., Saligram, B., Moses, M. E., Hatzakis, N. S., & Montoya, G. (2018). Conformational Activation Promotes CRISPR-Cas12a Catalysis and Resetting of the Endonuclease Activity. *Cell*, *175*(7), 1856-1871. https://doi.org/10.1016/j.cell.2018.10.045

Swarts, D. C., & Jinek, M. (2018). Cas9 versus Cas12a/Cpf1: Structure–function comparisons and implications for genome editing. *Wiley Interdisciplinary Reviews: RNA*, *9*(5), 1–19. https://doi.org/10.1002/wrna.1481

Tatu Ylönen (1995). *PuTTY - World's Most Popular Free SSH Client*. Retrieved on the 28th December of 2019 from https://www.ssh.com/ssh/putty

The Sherrill Group. (n.d.) *Computational Chemistry.* Retrieved on the 21st of October from http://vergil.chemistry.gatech.edu/courses/chem4681/background/node1.html

Tiwari, M. K., Singh, R., Singh, R. K., Kim, I. W., & Lee, J. K. (2012). Computational Approaches for Rational Design of Proteins With Novel Functionalities. *Computational and Structural Biotechnology Journal*, *2*(3). https://doi.org/10.5936/csbj.201209002

Turanli-Yildiz, B., Alkim, C., & Petek, Z. (2012). Protein Engineering Methods and Applications. *Protein Engineering*. https://doi.org/10.5772/27306

Vieira, D. S., Lourenzoni, M. R., Fuzo, C. A., Ward, R. J., & Degrève, L. (2012). Chapter: Structural Bioinformatics for Protein Engineering. *Innovations in Biotechnology*. 415-438 https://doi.org/10.5772/28658

Walker, R. & Steinbrecher, T. (2006) *AMBER ADVANCED TUTORIALS - TUTORIAL 3 - MM-PBSA*. Retrieved on the 12th march 2020 from http://AMBERmd.org/tutorials/advanced/tutorial3/index.htm

Wang, J. (2006) *GAFF*. Retrieved on the 26th October 2020 from http://ambermd.org/antechamber/gaff.html

Wang, M., Mao, Y., Lu, Y., Wang, Z., Tao, X., & Zhu, J. K. (2018). Multiplex gene editing in rice with simplified CRISPR-Cpf1 and CRISPR-Cas9 systems. *Journal of Integrative Plant Biology*, *60*(8), 626–631. https://doi.org/10.1111/jipb.12667

Wijma, H. J., & Janssen, D. B. (2013). Computational design gains momentum in enzyme catalysis engineering. *FEBS Journal*, *280*(13), 2948–2960. https://doi.org/10.1111/ febs.12324

Yamano, T., Nishimasu, H., Zetsche, B., Hirano, H., Slaymaker, I. M., Li, Y., Fedorova, I, Nakane, T., Makarova, K. S., Koonin, E. V., Ishitani, R., Zhang, F. & Nureki, O. (2016). Crystal Structure of Cpf1 in Complex with Guide RNA and Target DNA. *Cell*, *165*(4), 949–962. https://doi.org/10.1016/j.cell.2016.04.003

Yang, W., Lee, J. Y., & Nowotny, M. (2006). Making and Breaking Nucleic Acids: Two-Mg2+ Ion Catalysis and Substrate Specificity. *Molecular Cell 22 (1)*, 5-13. doi:10.1016/j.molcel.2006.03.013

Yin, H., Kauffman, K. J., & Anderson, D. G. (2017). Delivery technologies for genome editing. *Nature*, 1–13. https://doi.org/10.1038/nrd.2016.280

Zetsche, B., Gootenberg, J. S., Abudayyeh, O. O., Slaymaker, I. M., Makarova, K. S., Essletzbichler, P., Volz, S., Joung, J., Oost, J., Regev, A., Koonin, E. V. & Zhang, F. (2015). Cpf1 is a single RNA-guided endonuclease of a Class 2 CRISPR-Cas system. Cell, *163*(3), 759–771. https://doi.org/10.1016/j.cell.2015.09.038

Zhang, F., Wen, Y., & Guo, X. (2014). CRISPR/Cas9 for genome editing: Progress, implications and challenges. *Human Molecular Genetics*, *23*(R1), 40–46. https://doi.org/10.1093/hmg/ddu125

Zhang, Y., Long, C., Li, H., McAnally, J. R., Baskin, K. K., Shelton, J. M., Bassel-Duby, R., & Olson, E. N. (2017). CRISPR-Cpf1 correction of muscular dystrophy mutations in human cardiomyocytes and mice. *Science Advances*, *3*(4). https://doi.org/10.1126/sciadv.1602814

Anex I

List of Software

PuTTY

PuTTY software is a versatile terminal program available for Windows, Mac and Linux. It is the world's most popular free interface, as it supports several kind of connections between machines (Tatu Ylönen, 1995).

AMBER

AMBER (Assisted Model Building with Energy Refinement) consists on a collection of several programs - and empirical force fields used by them - which enable biomolecule dynamics simulations and analysis (Case, *et al.*, 2018). In order to do this, it uses CPUs (central processing unit) and GPUs (graphics processing unit). CPUs are fast and versatile, allowing the interaction between interdependent tasks, where what is processed in each core is relevant to the final project (intel, n.d.). It has few cores, contrasting with GPU composed by several hundreds. These last are the responsible for the Artificial Inteligence (AI) boom. GPUs have the capacity of processing the same task in several parallel cores at the same time - what makes them ideal for graphics and animations, where several activities - such as lighting and shapes' rendering - must occur simultaneously, henceforth its applications for videogames (Caulfield, 2009).

This work was developed using AMBER18 and the programs:

Leap - a platform which allows biomolecules modelling, it includes the programs tleap and xleap. In order to model the structures of interest - being those pre-existing structures or modified systems - it requires cartesian coordinates for each atom of interest within the system - aquired from a file with pdb format (that will also provide the topology of each atom) - and indications on the force fields applied to the system (Case, *et al*., 2018).

Sander - a tool used to optimize and equilibrate the system and to run the intended dynamics. The Molecular Dynamics simulations performed, generate different configurations of the structure in study, in accordance to Newtonian equations of motion. This program allows the development of simple simulations as well as more complex dynamics, which may include, for example, the introduction of constraints to the

structure. The dynamics generated can be saved regularly and posteriorly analyzed (Case, *et al*., 2018).

Pmemd (Particle Mesh Ewald Molecular Dynamics) - it can be considered a sander's optimization, as it allows the performance of faster simulations by using parallel scaling (Salomon-Ferrer, *et al.*, 2013).

Cpptraj - this program is used as a tool to analyze the dynamics performed. With this tool, it is possible to study particle trajectories, as well as atomic fluctuations and bond/angle/dihedral values, to give some examples (Case, *et al*., 2018).

MM-PBSA - this software is used to "calculate the free-energy change between two states" through Molecular Mechanics - Poisson Boltzmann Surface Area calculations. In the present case, it means that the energy of the system will be calculated before and after the binding of the protein and the nucleic acids. This is done by combining different snapshots of the structures for several conformations, both independently of a solvent model and solvation free components and considering an implicit solvent model (Case, *et al*., 2018).

UCSF Chimera

UCSF Chimera is "an Extensible Molecular Modelling System". This program can be used to interactively visualize and analyze molecular structures and its associated data, which may include sequence alignments and atom trajectories. It also allows the structures' modelling and was designed with extensibility as one of its goals. With Chimera, it is also possible to caption images and generate animations of the biomolecule of interest. (Pettersen, *et al.*, 2004)

VMD (Visual Molecular Dynamics)

As its name indicate, this program allows the visualization of a molecule and its dynamics. With VMD, it is possible to observe, as well as animate and analyze biomolecular systems - read from pdf files - through the use of three dimensional graphics and built-in scripts. This program also allows the generation of images and animations, being possible to edit each particle properties and visual characteristics - such as molecular representation, coloring style and transparency and materials, to give some examples. In this work, VMD was used to analyze the proteins' structure and its particles trajectory after each simulation (Humphrey, *et al*., 1996).

Anex II

Part1

Selection of atoms for the initial distance analysis for identification of which residues to mutate.

Residue - Atom	Nucleotide - Atom
ASN124 - HD22	DT10 - OP1
GLN125 - HB2	DT10 - OP1
GLN125 - H	DT10 - P
THR176 - HG22	DT10 - OP2
THR177 - HG23	DT10 - C7
LYS180 - HZ1	DT10 - C7
ASP615 - OD2	DT10 - O4
PRO670 - O	DA12 - O4’
PRO670 - HB2	DA12 - O3’
GLN704 - HE21	DA12 - O3’
LYS705 - HZ1	DA12 - OP2
LYS705 - HZ2	DA12 - C8
LYS705 - HZ3	DA12 - C3’
ARG800 - H	U15 - O3’
ARG800 - HD3	A16 - H5’’
ARG800 - H	A16 - OP1
ARG800 - O	A16 - OP2
ARG800 - HE	A16 - O3’
LYS798 - HZ2	U15 - OP1
LYS798 - HZ2	U15 - H5’’
LYS798 - HG2	U15 - H5’’
LYS798 - HZ1	G14 - H4’
LYS798 - HZ1	U13 - OP1
LYS798 - HZ3	U12 - OP1
LYS798 - HZ3	U13 - OP1

LYS798 - HZ1	G14 - O3’
LYS798 - HZ2	G14 - O3’
LYS791 - HE3	U15 - O4
LYS791 - HE2	U15 - H5
LYS791 - HB2	U15 - H5’
LYS791 - HE2	U15 - OP2
SER797 - HB2	G14 - H5’
SER797 - H	U15 - O5’
SER797 - HG	U15 - OP2
SER797 - HG	U15 - OP1
SER797 - HA	G14 - H5’
SER797 - HA	G14 - H4’
SER794 - HG	G14 - H4’
SER794 - HG	G14 - OP1
SER794 - HG	G14 - O5’
SER794 - OG	G14 - O5’
SER794 - OG	U13 - HO2’
ASN851 - OD1	U1312 - H5
ASN851 - HB3	U10 - OP2
ASN851 - HB3	U10 - OP1
ASN851 - HD21	U10 - HO2’
ASN851 - OD1	U10 - H3’
ASN851 - HB3	U10 - H3’
ASN851 - HA	U10 - H3’
ASN851 - OD1	U10 - H2’
ASN851 - OD1	U10 - HO2’
ASN851 - OD1	U10 - H6
ASN851 - HA	U10 - H5
ASN851 - HD22	U10 - H5’’
ASN851 - HD2	U10 - OP2
LYS852 - H	U10 - OP2
LYS852 - H	U10 - OP1
LYS852 - HB3	U10 - P

LYS852 - HD3	U10 - H5
LYS852 - HB3	U10 - H5
ASN853 - H	U10 - OP2
ASN853 - H	U10 - OP1
ASN853 - HB3	U10 - OP1
ASN853 - HB2	U10 - OP1
ASN853 - HD22	U10 - H5’
ASN856 - HD21	U12 - H5
ASN856 - HD21	U10 - H5’’
ASN856 - HD22	U10 - H5’’
ASN856 - HD22	U10 - OP1
ASN856 - HB2	U10 - OP1
ASN856 - HD21	U10 - H3’
ASN856 - HD21	G11 - OP1
ASN856 - OD1	G11 - OP1
LYS858 - HZ1	U12 - O4
LYS858 - HZ2	U12 - O4
LYS858 - HZ3	U12 - H5
LYS858 - HZ3	U12 - O3’
SER861 - HB3	U12 - H5
SER861 - HB3	U12 - O4
SER861 - HB3	U12 - C4
SER861 - HG	U10 - O2’
SER861 - HB2	U10 - HO2’
SER861 - HG	U10 - HO’
SER861 - HG	U10 - H3’
VAL862 - O	U12 - H3
VAL862 - H	U12 - O4
PHE863 - HE1	U12 - H1’
PHE863 - HZ	U12 - H5
PHE863 - HE1	U12 - O4’
PHE863 - HA	U12 - O2
PHE863 - HZ	U10 - O2

PHE863 - HZ	U10 - O2’
PHE863 - HE2	U10 - HO2’
TYR865 - HB3	U13 - N3
TYR865 - HB2	U13 - N3
TYR865 - HD1	U13 - N3
TYR865 - HE1	U13 - O4’
TYR865 - HE1	U13 - O2
TYR865 - OH	U12 - HO2’
TYR865 - OH	U12 - O2’
TYR865 - OH	G14 - N7
TYR865 - OH	U13 - OP2
TYR865 - HH	U12 - H1’
ASP866 - O	U13 - O4
ASP866 - O	U13 - H3

Part2

Atoms selected for distance analysis

Wild type structure

Designation	Residue - Atom	Nucleotide - Atom
D1	ASP124 - HD21	DT10 - OP1
D2	ASP124 - HD22	DT10 - OP1
D3	ASP124 - HD21	DT10 - O3’
D4	ASP124 - HD22	DT10 - O3’
D5	ASP124 - HD21	DT11 - OP1
D6	ASP124 - HD22	DT11 - OP1
D7	ASP124 - HD21	DT11 - OP2
D8	ASP124 - HD22	DT11 - OP2
D9	ASP856 - HD21	U10 - OP1
D10	ASP856 - HD22	U10 - OP1
D11	ASP856 - HD21	U10 - O5’
D12	ASP856 - HD22	U10 - O5’
D13	ASP856 - HD21	G11 - OP1
D14	ASP856 - HD22	G11 - OP1
D15	VAL862 - O	U12 - H3
D16	VAL862 - H	U12 - O4

N124K mutant

Designation	Residue - Atom	Nucleotide - Atom
D1	LYS124 - HZ1	DT11 - OP2
D2	LYS124 - HZ2	DT11 - OP2
D3	LYS124 - HZ3	DT11 - OP2
D4	LYS124 - HZ1	DT11 - OP1

D5	LYS124 - HZ2	DT11 - OP1
D6	LYS124 - HZ3	DT11 - OP1
D7	LYS124 - HZ1	DT10 - O3’
D8	LYS124 - HZ2	DT10 - O3’
D9	LYS124 - HZ3	DT10 - O3’

N856K mutant

Designation	Residue - Atom	Nucleotide - Atom
D1	LYS856 - HZ1	G11 - OP1
D2	LYS856 - HZ2	G11 - OP1
D3	LYS856 - HZ3	G11 - OP1
D4	LYS856 - HZ1	G11 - OP2
D5	LYS856 - HZ2	G11 - OP2
D6	LYS856 - HZ3	G11 - OP2
D7	LYS856 - HZ1	G11 - O5’
D8	LYS856 - HZ2	G11 - O5’
D9	LYS856 - HZ3	G11 - O5’

V862K mutant

Designation	Residue - Atom	Nucleotide - Atom
D1	LYS862 - O	U12 - H3
D2	LYS862 - H	U12 - O4

N124K_ N856K_ V862K mutant

Designation	Residue - Atom	Nucleotide - Atom
D1	LYS124 - HZ1	DT11 - OP1
D2	LYS124 - HZ2	DT11 - OP1
D3	LYS124 - HZ3	DT11 - OP1
D4	LYS124 - HZ1	DT11 - OP2

D5	LYS124 - HZ2	DT11 - OP2
D6	LYS124 - HZ3	DT11 - OP2
D7	LYS124 - HZ1	DT10 - O3’
D8	LYS124 - HZ2	DT10 - O3’
D9	LYS124 - HZ3	DT10 - O3’
D10	LYS856 - HZ1	U10 - O2’
D11	LYS856 - HZ2	U10 - O2’
D12	LYS856 - HZ3	U10 - O2’
D13	LYS856 - HZ1	U10 - O3’
D14	LYS856 - HZ2	U10 - O3’
D15	LYS856 - HZ3	U10 - O3’
D16	LYS856 - HZ1	G11 - OP1
D17	LYS856 - HZ2	G11 - OP1
D18	LYS856 - HZ3	G11 - OP1
D19	LYS856 - HZ1	G11 - OP2
D20	LYS856 - HZ2	G11 - OP2
D21	LYS856 - HZ3	G11 - OP2
D22	LYS856 - HZ1	U12 - O4
D23	LYS856 - HZ2	U12 - O4
D24	LYS856 - HZ3	U12 - O4
D25	LYS862 - O	U12 - H3
D26	LYS862 - H	U12 - O4

Angles selected for distance analysis

Wild type structure

Designation	Residue - Atom	Residue - Atom	Nucleotide - Atom
A1	ASP124 - ND2	ASP124 - HD21	DT10 - OP1
A2	ASP124 - ND2	ASP124 - HD22	DT10 - OP1
A3	ASP124 - ND2	ASP124 - HD21	DT10 - O3’

A4	ASP124 - ND2	ASP124 - HD22	DT10 - O3’
A5	ASP124 - ND2	ASP124 - HD21	DT11 - OP1
A6	ASP124 - ND2	ASP124 - HD22	DT11 - OP1
A7	ASP124 - ND2	ASP124 - HD21	DT11 - OP2
A8	ASP124 - ND2	ASP124 - HD22	DT11 - OP2
A9	ASP856 - ND2	ASP856 - HD21	U10 - OP1
A10	ASP856 - ND2	ASP856 - HD22	U10 - OP1
A11	ASP856 - ND2	ASP856 - HD21	U10 - O5’
A12	ASP856 - ND2	ASP856 - HD22	U10 - O5’
A13	ASP856 - ND2	ASP856 - HD21	G11 - OP1
A14	ASP856 - ND2	ASP856 - HD22	G11 - OP1
A15	VAL862 - O	U12 - H3	U12 - N3
A16	VAL862 - N	VAL862 - H	U12 - O4

N124K mutant

Designation	Residue - Atom	Residue - Atom	Nucleotide - Atom
A1	LYS124 - NZ	LYS124 - HZ1	DT11 - OP2
A2	LYS124 - NZ	LYS124 - HZ2	DT11 - OP2
A3	LYS124 - NZ	LYS124 - HZ3	DT11 - OP2
A4	LYS124 - NZ	LYS124 - HZ1	DT11 - OP1
A5	LYS124 - NZ	LYS124 - HZ2	DT11 - OP1
A6	LYS124 - NZ	LYS124 - HZ3	DT11 - OP1
A7	LYS124 - NZ	LYS124 - HZ1	DT10 - O3’
A8	LYS124 - NZ	LYS124 - HZ2	DT10 - O3’
A9	LYS124 - NZ	LYS124 - HZ3	DT10 - O3’

N856K mutant

Designation	Residue - Atom	Residue - Atom	Nucleotide - Atom
A1	LYS856 - NZ	LYS856 - HZ1	G11 - OP1
A2	LYS856 - NZ	LYS856 - HZ2	G11 - OP1

A3	LYS856 - NZ	LYS856 - HZ3	G11 - OP1
A4	LYS856 - NZ	LYS856 - HZ1	G11 - OP2
A5	LYS856 - NZ	LYS856 - HZ2	G11 - OP2
A6	LYS856 - NZ	LYS856 - HZ3	G11 - OP2
A7	LYS856 - NZ	LYS856 - HZ1	G11 - O5’
A8	LYS856 - NZ	LYS856 - HZ2	G11 - O5’
A9	LYS856 - NZ	LYS856 - HZ3	G11 - O5’

V862K mutant

Designation	Residue - Atom	Residue - Atom	Nucleotide - Atom
A25	LYS862 - O	U12 - H3	U12 - N3
A26	LYS862 - N	LYS862 - H	U12 - O4

N124K_ N856K_ V862K mutant

Designation	Residue - Atom	Residue - Atom	Nucleotide - Atom
A1	LYS124 - NZ	LYS124 - HZ1	DT11 - OP1
A2	LYS124 - NZ	LYS124 - HZ2	DT11 - OP1
A3	LYS124 - NZ	LYS124 - HZ3	DT11 - OP1
A4	LYS124 - NZ	LYS124 - HZ1	DT11 - OP2
A5	LYS124 - NZ	LYS124 - HZ2	DT11 - OP2
A6	LYS124 - NZ	LYS124 - HZ3	DT11 - OP2
A7	LYS124 - NZ	LYS124 - HZ1	DT10 - O3’
A8	LYS124 - NZ	LYS124 - HZ2	DT10 - O3’
A9	LYS124 - NZ	LYS124 - HZ3	DT10 - O3’
A10	LYS856 - NZ	LYS856 - HZ1	U10 - O2’
A11	LYS856 - NZ	LYS856 - HZ2	U10 - O2’
A12	LYS856 - NZ	LYS856 - HZ3	U10 - O2’
A13	LYS856 - NZ	LYS856 - HZ1	U10 - O3’
A14	LYS856 - NZ	LYS856 - HZ2	U10 - O3’
A15	LYS856 - NZ	LYS856 - HZ3	U10 - O3’

A16	LYS856 - NZ	LYS856 - HZ1	G11 - OP1
A17	LYS856 - NZ	LYS856 - HZ2	G11 - OP1
A18	LYS856 - NZ	LYS856 - HZ3	G11 - OP1
A19	LYS856 - NZ	LYS856 - HZ1	G11 - OP2
A20	LYS856 - NZ	LYS856 - HZ2	G11 - OP2
A21	LYS856 - NZ	LYS856 - HZ3	G11 - OP2
A22	LYS856 - NZ	LYS856 - HZ1	U12 - O4
A23	LYS856 - NZ	LYS856 - HZ2	U12 - O4
A24	LYS856 - NZ	LYS856 - HZ3	U12 - O4
A25	LYS862 - O	U12 - H3	U12 - N3
A26	LYS862 - N	LYS862 - H	U12 - O4